Decision Support Systems xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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

A Tabu search heuristic for smoke term curation in safety defect discovery

David M. Goldberg *, Alan S. Abrahams

Department of Business Information Technology, Pamplin College of Business, Virginia Tech, Pamplin Hall, Suite 1007, 880 West Campus Drive, Blacksburg, VA 24061, United States

ARTICLE INFO

Article history:
Received 29 April 2017
Received in revised form 20 October 2017
Accepted 20 October 2017
Available online xxxx

Keywords:
Text mining
Online reviews
Tabu search
Heuristics
Defects
Business intelligence

ABSTRACT

The ability to detect and rapidly respond to the presence of safety defects is vital to firms and to regulatory agencies. In this paper, we employ a text mining methodology to generate industry-specific "smoke terms" for identifying these defects in the countertop appliances and over-the-counter medicine industries. Building upon prior work, we propose several methodological improvements to enhance the precision of our industry-specific terms. First, we replace the subjective manual curation of these terms with an automated Tabu search algorithm, which provides a statistically significant improvement over a sample of human-curated lists. Contrary to the assumptions of prior work, we find that shorter, targeted smoke term lists produce superior precision. Second, we incorporate non-textual review features to enhance the performance of these smoke term lists. In total, we find greater than a twofold improvement over typical human-curated lists. As safety surveillance is vital across industries, our method has great potential to assist firms and regulatory agencies in identifying and responding quickly to safety defects.

1. Introduction

Product defects are enormous concerns for manufacturers across industries. The costs to firms of recent recalls have reached billions of dollars for defects to single SKUs of products; in the electronics industry, Samsung's estimated loss from the recall of its Note 7 phone was $5.3 billion [1], while the recall of Takata airbags in the automotive industry was estimated to cost the firm up to $24 billion [2]. Safety and performance defects are both concerning for firms, but safety defects often invoke harsher responses due to the capacity for causing bodily harm to consumers, and unlike performance defects, they may result in recalls issued by the Consumer Product Safety Commission (CPSC), Food and Drug Administration (FDA), or other federal agencies. Furthermore, safety defects are concerning to firms because associated recalls not only result in explicit costs to repair damages, but implicit costs are also likely because news stories on safety defects in a firm's products tend to damage goodwill [3].

From the perspective of manufacturers, the task of identifying and responding to safety defects is complex. Manufacturers may conduct testing on their products in quality control departments to prevent some safety defects before products reach consumers. In addition, manufacturers may review warranty claims for their products to understand the causes of defects. However, the conditions of consumers' uses of products are difficult to reproduce exactly in quality control testing [3], and the prevalence of product recalls at over $1 trillion of total costs in the United States each year [4] indicates that detection of safety defects after products reach the mass market is paramount. To this end, many firms and regulatory agencies have recently begun employing teams to seek out discussions of safety defects online. As the Internet has provided a vibrant medium for the discussion of products across the globe, discussion forums and product reviews have provided a massive new data source. However, despite the immense value of the volume of data available, the unstructured nature of textual data poses challenges for the detection of safety defects, as it is unrealistic for human readers to keep up with the pace of new online content [5]. Firms may be pleased that a minority of online reviews refers to safety defects, but this facet makes identifying and prioritizing the set of reviews actually referring to those defects a difficult task. Furthermore, there is substantial evidence that consumers read online reviews to inform their purchasing decisions [6–8], so minimizing the extent to which online reviews represent defect-laden feedback about products is a substantial concern for manufacturers.

Only recently has research on automated detection of defects started to take shape in the literature. The key work by Abrahams et al. [5,9,10] establishes a framework by which defects may be detected in these online media. Rather than relying on traditional automated sentiment analysis dictionaries, the methodology proposes creating industry-specific lists of "smoke terms", or terms particularly associated with defects in that industry [5,9,10]. Beyond this initial work in the automotive industry, further research by Winkler et al. [11], Law et al. [12], and Adams et al. [13] has applied these techniques in the toy, dishwasher, and joint/muscle treatment industries respectively, to great effect. Although automated techniques are now applied as a standard component of defect discovery analysis, humans perform the "curation" process, or the choosing of the final terms. Information retrieval

* Corresponding author.
E-mail address: goldberg@vt.edu (D.M. Goldberg).

https://doi.org/10.1016/j.dss.2017.10.012
0167-9236/

techniques such as those proposed in Fan et al. [14] generate a ranked list of term relevance based on a training sample; from that initial list of terms, human judgment is employed to filter relevant from irrelevant terms for inclusion in the final smoke term lists [5,9–13]. This approach presents two key limitations. First, due to the inherent subjectivity of determining which terms ought to be considered relevant, this procedure introduces the possibility of substantial variance in performance between the lists generated by different individuals. To the best of our knowledge, the literature has not yet studied the variability in performance across these lists. However, the possibility of such variability is an enormous potential problem, as curating high performing lists should allow firms and regulatory agencies to identify and respond to defects in an expedient manner. Second, the manual curation of these smoke terms represents an additional labor requirement for organizations. These organizations may be unsure how best to curate these lists, and they may also lack available labor to devote to the task.

A further limitation of the status quo approach is that it focuses purely on textual characteristics of online media, but it does not incorporate other characteristics of these media. Of course, the textual data contained in online media may provide the clearest reference to the presence of a defect; however, further attributes of the online media may be useful means to verify or augment these textual characteristics.

In this work, we propose to build upon contemporary literature in defect discovery by addressing the aforementioned limitations in the smoke term methodology. We obtained a large sample of online reviews from the countertop appliances and over-the-counter (OTC) medicine industries for study in this paper. The countertop appliances industry has received great attention in recent times for safety defects, including a wide range of products recalled due to concerns of catching fire [15]. Additionally, appliances such as blenders contain fast-moving parts that may detach and become hazardous to bystanders; in a recent high profile story, several Cuisinart appliances were recalled by the CPSC [16]. Recalls in the OTC medicine industry are also problematic, such as a 2016 nationwide recall of potentially harmful children's medications [17]. As such, analysis of these industries ought to provide a ripe data source for our analysis. To establish a baseline of human performance at the task of smoke term curation, we asked an array of human participants to perform the task on our datasets, and we observed wide-ranging results. We propose a Tabu search algorithm for use in smoke term curation, which we find offers a statistically significant improvement in performance relative to human-curated smoke term lists. Although prior research has generally assumed that the inclusion of many smoke terms improves precision [9,11,13], we actually find that shorter and more targeted lists often offer superior performance. Additionally, we propose a scheme of augmenting both human-curated and machine-curated lists, treating star ratings as an interaction term and causing negative reviews in which star ratings are aligned with textual content to score particularly high values. We find that this method produces further statistically significant improvement upon both human-curated and machine-curated smoke term lists.

The remainder of this paper is structured as follows. First, we provide a comprehensive literature review on online reviews, text and sentiment analyses, and smoke term curation to motivate the value of an automated technique to improve curation. We describe the contributions of this work as well as the key research questions that we seek to address. We then lay out the new methodology that we propose in contrast to the methodology of prior work. Using our datasets, we provide results contrasting the performance of our technique to previous defect detection techniques. We note several of the potential limitations of our technique. Finally, we conclude our paper and present an overview of its implications as well as some opportunities for future work.

2. Literature review

In this section, we provide a review of related work on online reviews, text and sentiment analyses, and smoke term curation. We discuss the areas of coverage for prior work as well as limitations and unanswered questions. In particular, we conclude the section by discussing the subjective manner in which manual smoke term curation occurs, and we elaborate upon the possibility of improving this methodology.

2.1. Online reviews

As the availability of the Internet has expanded worldwide, online word-of-mouth (WOM) communication has been recognized as an important indicator of consumer opinion for products, and it serves as a window into product sales and product quality. WOM communication refers to the informal interchange of information by users concerning the characteristics, desirability, and use of products [6]. WOM communication in online reviews includes vital information on those consumers' perceptions of product quality [7], and these reviews have further impacts upon future consumption of those products by other consumers reading the reviews [6]. Some of the largest online retailers, such as Amazon, Best Buy, and Target, provide online review platforms for consumers to share their experiences with products, and these platforms have become staples of online shopping experiences.

Consumers treat online reviews as a key source of information when learning about products online. A survey by BrightLocal [8] found that 91% of consumers read online reviews to better understand the quality of products they are interested in before purchase, and 84% of consumers trust online reviews equivalently to personal recommendations. Research has indicated a relationship between online reviews and the sales of reviewed products [7]. Chevalier and Mayzlin [6] found evidence that the mean star rating in product reviews was positively related with the subsequent sales of associated products, while Duan et al. [18] found that the volume of reviews for products is positively related with subsequent sales, possibly serving as a proxy for product popularity. Importantly, multiple aspects of online reviews reflect consumers' opinions. For example, Mudambi et al. [19] discuss the potential for misalignment between the textual content of a consumer's review and associated star ratings. As such, consideration of both textual and non-textual aspects of reviews may offer essential insights.

Online reviews not only provide enormous volumes of data about products, but they also provide data from a wide array of customers in an accessible format for researchers and practitioners alike. The diversity of users for each product also ensures a diversity of uses for each product, and, as such, safety defects may only be detectable by some parts of the customer base. Therefore, the enormous volume of customer experiences provided by online reviews serves as an invaluable tool in defect discovery studies.

2.2. Text and sentiment analyses

Due to the spread of Internet connectivity around the world, firms are now faced with a plethora of unstructured data in textual format. As such, text and social media analyses, algorithms for extracting insights from this type of data, have proved to be key areas in Big Data analytics [18,20]. Researchers extract text from online sources, such as product reviews [5–7,9–13] and social media [21] to support decision-making.

Sentiment analysis refers to a broad family of natural language processing techniques employed to assess the type(s) and amount of emotion expressed in text. Frequently, sentiment analysis involves the use of sentiment dictionaries in which words are associated with quantitative valence scores. Examples of such sentiment dictionaries include AFINN [22], ANEW [23], and the Harvard General Inquirer [24]. Some sentiment analysis techniques such as SentiStrength [25] attempt to augment these analyses by incorporating context of surrounding words and phrases.

Researchers have employed sentiment analysis extensively to understand online product reviews and discussion forums, as a consumer's textual valence with respect to a product serves as an important indicator of their opinion [26]. Tang et al. [27] provide a comprehensive overview of prior sentiment analysis literature in online reviews. Sentiment

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

analysis has even been used to predict stock performance at the firm [20] and market levels [21].

Due to its ability to distinguish negative from positive opinions, sentiment analysis has been employed to detect product safety concerns [28,29]. Indeed, online comments invoking especially negative sentiment logically would seem more likely to refer to safety concerns than online reviews invoking positive sentiment. While on the surface, sentiment analysis seems to be a viable method of detecting safety defects in online reviews, researchers have also pointed out several flaws in these methods [12,13]. First, sentiment dictionaries typically depend on quantifying the emotive valence of specific words, but online reviews are rife with exceptions to these rules. For example, although the word "problem" may be classified as negative by most sentiment dictionaries, online reviews may contain statements such as, "the set-up for this product was no problem". Second, delineating the specific type of complaint of interest may be challenging with traditional sentiment dictionaries. While organizations may be very concerned with safety defects, sentiment analysis may capture a great deal of performance-related issues instead, as negative sentiment does not distinguish between types of customer dissatisfaction with products. Finally, while sentiment dictionaries capture a great deal of negative valence terms, they may fail to capture domain-specific concerns that human readers would quickly identify as noteworthy. For example, a review of a furniture product may contain statements such as, "the dresser appeared to teeter", which does not contain any largely emotive words; yet, the potential instability of the furniture product obviously illustrates an enormous potential safety concern.

Although sentiment analyses clearly have a valuable place in investigating online reviews, the aforementioned limitations represent substantial concerns that they may not be the most effective choice for safety defect detection. Thus, an approach specifically targeting these defects ought to be more effective.

2.3. Smoke term curation

As opposed to broader sentiment analyses, the literature has found industry-specific evidence that consumers describe performance and safety defects in products using particular words and phrases (*n*-grams) corresponding to the nature of the products in that industry [5,9–13]. For example, although the term "airbag" may not be associated with negative sentiment according to most sentiment dictionaries, it likely reflects a safety concern in the context of a consumer's online posting about a vehicle. A substantial stream of research in defect discovery focuses on this issue, generating industry-specific lists of "smoke terms" designed to identify defect-related language [5,9–13].

A critical stage of the smoke term curation process involves using information retrieval techniques to rate the relevance of terms in a corpus. Typically, researchers delineate a training sample with which to evaluate the relevance of terms, and the precision of these terms is evaluated in a separate holdout sample [5,9]. The literature describes several methods for rating the prevalence of these terms in the training sample. Robertson's Selection Value (RSV) is a method based on probability theory that estimates the likelihood of the presence of each term given that a document is relevant [30]. Fan et al. [14]'s pivotal work proposes several further innovative strategies for evaluating term relevance. First, Fan et al. [14] propose the Relevance Correlation Value (RCV), based on the Vector Space model [31], for defining a term's relevance based upon the number of times it appears in training documents classified as relevant. Fan et al. [14] also propose an adaptation upon Ng et al. [32]'s Correlation Coefficient (CC) metric that uses the X^2 distribution to assess whether two categorical variables, the occurrences of a word and the relevance statuses of documents, occur independently. Each of these relevance scores has received substantial attention in experiments in the research community; RSV and RCV metrics were employed in Abrahams et al. [9,10], although more recent works on defect discovery have utilized the CC score extensively and observed excellent performance [11–13]. The scores assigned by these techniques serve as weights upon each term in the final analysis: terms with greater scores have greater weights in marking a document's language as referring to safety defects. A detailed description of this process and an example dataset may be found in the Online Supplement.

After initially using the aforementioned information retrieval techniques to obtain a relevance score for each of the terms in a corpus, researchers and practitioners are tasked with curating a smoke term list. Although the terms scoring the highest relevance values are generally believed to be most suitable for smoke term lists, many of the comparatively less relevant terms must be removed from the lists to ensure that defects are distinctly identified [5,9–13]. Indeed, research has observed a decline in the quality of terms after those ranked in the top few hundred [10,11]. Researchers often set arbitrary cutoffs for the number of terms to include in final smoke term lists or minimum relevance scores for inclusion [11,12]; in addition, smoke term lists are manually filtered to retain only the terms intuitively believed to provide the greatest precision [5,9–13]. The literature has provided several rationales for the removal of terms from this initial list but acknowledges that this process is substantially subjective [5,9–13].

First, it is typical for researchers to remove common English words or "stop words" such as "a", "an", and "the" that may be highly prevalent in defect-tagged reviews [5,9–13]. These words may be highly prevalent in reviews referring to safety defects, but they do not indicate safety defects in and of themselves, presenting a chance for false positives if included. Second, researchers frequently remove common product or brand terms that are highly prevalent in defect-tagged reviews [5,9–13]. For example, in the toy industry, the terms "doll" or "helicopter" may merely identify the types of products available rather than the defectiveness of those products, or brand names like "Hasbro" or "Mattel" may be prevalent due to brand popularity biases [11]. Of course, the removal of these terms is potentially risky if they refer to elements of products that are actually defective. For example, a brand name may be worthy of inclusion if most of its products are actually associated with defects. Third, researchers may wish to remove sub-product terms that cause many predicted-hazard false-positives [5,9–13]. For example, the terms "arm", "leg", and "hair" often refer to specific parts of dolls in the toy industry and are mentioned regularly in non-hazard reviews, whereas in other industries these words may be more likely to refer to injured body parts of the consumer [11]. Conversely, some researchers include these terms as references to specific defective components of products. For example, in the automotive industry, references to "airbags" likely refer to the defective nature of those parts [5,10]. Finally, researchers may remove spuriously prevalent words subjectively identified as irrelevant [5,9–13].

Of course, determining whether and the extent to apply these rules is subjective. The efficacy of the resultant smoke term list at identifying defects greatly depends upon the specific terms included in that list. Excluding a relevant term may result in the final smoke term list entirely neglecting an important category of safety defects. On the other hand, including an irrelevant term may introduce false positives. Currently, the efficacy of various smoke term choices is an open research question requiring substantially more study.

3. Research questions and contribution

In this paper, we address three key research questions. First, to what extent does the performance of smoke term lists vary across human curators? Second, to what extent do heuristic methods for smoke term curation improve upon the performance of human-curated lists? Third, to what extent does the inclusion of non-textual review data improve the performance of smoke term lists?

We make three key contributions in this paper. First, to the best of our knowledge, we provide the first study comparing the efficacy of human-curated smoke term lists across individuals. Although smoke term lists have been assumed to be well curated in prior work utilizing this methodology [5,9–13], the issue of the variability of the efficacy of these lists has not yet been analyzed in the literature. We provide details

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

on the extent of this issue and the level of performance that may be expected in smoke term lists. Second, we make substantial methodological enhancements to the existing literature that result in statistically significant improvements in the performance of smoke term lists. We utilize a Tabu search algorithm to automate the process of smoke term curation, which results in lists that outperform all human-curated lists. Contrary to the principles assumed in prior work [9,11,13], we actually find that relatively short lists often offer superior performance. Not only does this innovation improve upon the performance of lists generated in the status quo, but it also alleviates a labor requirement for firms and regulatory agencies by automating the previously subjective process. We provide the first incorporation of non-textual characteristics in smoke term lists, and we find that the utilization of star ratings to augment smoke term scores results in a statistically significant improvement in performance for both human-curated and machine-curated smoke term lists. This improved methodology may be applied within any industry for detecting defects, easing the process of rapidly responding to safety hazards. Third, we define a new class of smoke terms for the countertop appliances and OTC medicine industries, which may be applied immediately for the detection of defects. Firms and regulatory agencies may make use of these terms for identifying and responding to defects as quickly as possible.

4. Methodology

4.1. Aims of the technique

In defect detection, it is of paramount importance for firms and regulators alike that reviews clearly indicating defective products are scored as such by the employed text mining algorithm(s). Due to constraints on time and resource capabilities and the volume of data that appears online every day, it is unreasonable to expect firms or regulatory agencies to read every review of a product to analyze it for potential defects [5]. Indeed, the purpose of the smoke term methodology is to rate every review in a corpus for the extent to which it appears to contain defect-related language. As such, practitioners are not tasked with reading every review, but only the top portion of reviews. For this reason, recent studies implementing this methodology have evaluated the efficacy of their techniques by focusing on the precision obtained in the top *N*-ranked reviews as scored by the algorithm. The portion of reviews feasible to read of course depends upon the needs of the firm or regulatory agency in addition to the specific industry and the volume of reviews. However, in prior research, focus on the top 100-ranked [11] or 200-ranked reviews [12,13] is common. In practice, these top reviews represent the area of focus for the industry; indeed, laboriously reading every review of a product would be contrary to the purpose of such a scoring algorithm.

For the purpose of our technique, after ranking the corpus of reviews using a smoke term list from most relevant to least relevant, we then assess performance using the number of defects found in the top 50-ranked reviews, top 100-ranked reviews, and top 200-ranked reviews. Using this spread of performance metrics, we aim to show that our method provides excellent performance regardless of the chosen cutoff. In addition to these metrics, we provide Receiver Operating Characteristic (ROC) curves to observe the performance at arbitrary cutoffs.

4.2. Dataset and data coding

We chose Amazon.com, the world's largest e-commerce retailer and a substantial review platform, as the data source for this project [33]. In collaboration with a large manufacturer of countertop appliances with over $500 million in annual revenue, we randomly chose 100,000 countertop appliance reviews from Amazon for use in our study. Additionally, we randomly chose 12,400 over-the-counter (OTC) medicine reviews from Amazon for use in our study. These OTC medicines included allergy medicine, cough syrups, acetaminophen (pain relief), antacids, and digestion aids. We created a scheme of coding these reviews into two mutually exclusive categories: "safety defect" and "no safety defect" [5]. In the following, we describe the delineation between these two classes of reviews.

1) "Safety defects" refer to reviews that indicate a serious problem or malfunction in the functionality of a product that has caused or has the potential to cause bodily harm or property damage. Examples include electrical problems, smoke emission, or unsafe spillage of hot water from countertop appliances. For OTC medicine, examples include dangerous side effects, such as seizures or hallucinations. The following is an example of a customer review tagged as referring to a safety defect.

"*Leaks from coffee reservoir floor after 3 months -- coffee/water actually pours out from underneath machine. No gasket, just molded plastic which means it can't be fixed. Unfortunate as it makes a great cup of coffee. Note that this is a well-known problem with this coffeemaker…*"

2) "No safety defects" refer to reviews that contain other information and that do not refer to a specific safety-related problem. Positive product reviews and general comments are examples of "no safety defect" reviews, but negative reviews referring to performance concerns with products also fall into this category. Non-serious product malfunctions that result in poor or no functionality but that do not threaten human health or property are instances of "no safety defect". The following is an example of a review tagged as not referring to a safety defect.

"*This coffee grinder worked OK initially but after 6-7 moths, it started to stop grinding after 10–20 seconds. And after 9 months, it stopped working, so it went bad gradually.*"

In total, 545 undergraduate business students trained in quality management participated in the task of tagging the reviews as "safety defect" or "no safety defect". 442 students participated in the countertop appliances tagging project, and 103 students participated in the OTC medicine tagging project. In addition, a representative from the manufacturing firm with which we collaborated tagged a segment of 238 countertop appliance reviews as an "authority tagger" so that the student tags could be validated, and the lead researcher tagged 300 OTC medicine reviews for this purpose. In all cases, the reviews presented to taggers were randomly selected. Due to random presentation, some reviews were tagged by multiple taggers, and not all reviews were tagged. The taggers generated a total of 88,485 tags across 83,944 countertop appliances reviews and 13,794 tags across 10,874 OTC medicine reviews. In total, 4142 countertop appliance reviews were tagged multiple times, and amongst these instances, taggers were unanimous 3986 times (96.2% of cases) and disagreed just 156 times (3.6% of cases). Additionally, 2157 OTC medicine reviews were tagged multiple times, for which taggers were unanimous 2110 times (97.8% of cases) and disagreed just 47 times (2.2% of cases). In these cases of tagger disagreement, we assigned final designations to reviews with a "most conservative" rule: these reviews were each classified as safety defects [11]. As the cost of a false negative is especially high in safety surveillance, we opt to assume that these reviews reflect safety concerns. Due to random presentation, the tags of the authority tagger can be compared to the tags of other taggers to calculate inter-rater agreement statistics. For countertop appliances, we observed 93.3% tagging agreement between the authority tagger and other taggers and a Cohen's κ [34] value of 0.867. For OTC medicine, we observed 91.3% agreement between the authority tagger and other taggers and a Cohen's κ [34] value of 0.827. Landis and Koch [35] rate agreement in this range as "almost perfect", while Fleiss et al. [36] rate agreement in this range as "excellent".

In Fig. 1, we provide an overview of the methodological work to be performed in the remainder of the section. For each industry, we separate the set of tagged reviews into three approximately equally sized portions: (A) a training set, (B) a curation set, and (C) a holdout set.

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

Using the training set, we perform process (1), a ranking of the *n*-grams in the training set by relevance in safety defect-tagged reviews as measured by the CC score [14]. In turn, this process informs process (2), the curation of smoke terms from the initial list provided in the previous step. Following the examples of prior research [5,9,13], we select the top 200 unigrams, top 200 bigrams, and top 200 trigrams as manageable supersets. From each of these supersets, we use the Tabu search to obtain final unigram, bigram, and trigram smoke term lists that maximize precision in the curation set. We recruited a sample of individuals to perform manual curation upon the top 200 *n*-grams in each superset, and we compare this process to an automated Tabu search. These sets of smoke terms generated in (2) are used in (3) to compare the efficacy of human-curated versus machine-curated smoke term lists. As baseline comparisons, we also show the performance of common sentiment dictionaries and of random chance, or the rate of defect detection when sorting through the reviews randomly. Additionally, we show the effect of including star ratings in these evaluations as a method of boosting the scores of reviews believed to refer to safety defects and ameliorating false positives.

4.3. Initial smoke term delineation

We utilize information retrieval techniques to create an initial ranking of terms by relevance in safety defect-tagged reviews. Although the literature has acknowledged several different review scoring methods for this application, recent research on defect discovery [11–13] has found the best performance when using the CC score proposed by Fan et al. [14]. As such, we utilize this method to quantify relevance of each *n*-gram in safety defect-tagged reviews. In later sections, we employ the top 200-ranked terms for each of unigrams, bigrams and trigrams. Each term in the training set is ranked by the CC score, and prior research has found that the quality and relevance of the terms tends to attenuate as CC score diminishes, and term relevance seems to most severely attenuate for terms beyond the threshold of 200 terms with highest CC score [9,12,13]. These scores serve as weights upon each term in the final analysis, as terms with greater scores have greater impact in marking a review's safety defect-related language. A detailed description of this process and an example dataset may be found in the Online Supplement. For robustness, we also ran our Tabu search algorithm when considering several other thresholds, and we observed that the algorithm never recommended any of the terms ranked beyond the top 200 by the CC score.

In total, the training set of 27,981 countertop appliances reviews contained 29,281 unique unigrams, 448,890 unique bigrams, and 1,373,640 unique trigrams. The training set of 3624 OTC medicine reviews contained 9709 unique unigrams, 84,824 unique bigrams, and 163,732 unique trigrams. The vast number of *n*-grams in these datasets of online reviews illustrates the value of the information retrieval techniques. Without using such techniques to provide an initial set of rankings for *n*-grams with which to narrow down the term relevance, there would be far too many terms for humans or algorithms to generate smoke term lists in a reasonable timeframe. The top 10-ranked unigrams, bigrams, and trigrams by relevance (CC score) in each set of safety reviews are listed in Table 1.

4.4. Human smoke term curation

To assess human performance in smoke term curation, we recruited a sample of human participants to establish a baseline. We created a survey in which participants were able to curate their own smoke term lists given an initial set of the top 200 unigrams, the top 200 bigrams, and the top 200 trigrams as delineated by the CC score metric [14]. We then solicited participation for our survey using Amazon Mechanical Turk, which offers access to a global marketplace of over 500,000 workers to perform "human intelligence tasks" drawing from 190 countries. Relative to a choice of student participants, those participants from Amazon Mechanical Turk should offer a diverse set of backgrounds and experiences with which to inform their curation processes. Participants were provided with the initial lists of the top 200 unigram, top 200 bigram, and top 200 trigram smoke terms generated by the CC score [14] and were asked to identify from these lists the terms that they believed would delineate reviews referring to safety defects from other reviews. This configuration ensures conformity with the conditions of the Tabu search experiments as well as with prior work [5,9–13]. To avoid biasing the curation process, participants were not advised further as to which types of terms they ought to choose or the number of terms that they ought to choose. In total, we received 102 usable responses from Amazon Mechanical Turk for the countertop appliances industry and 139 usable responses for the OTC medicine industry. In the former sample, 59 respondents identified as male, while 43 respondents identified as female. The respondents ranged in age from 21 to 67; the average age was 34.6, and the standard deviation was 11.1. In the latter sample, 79 respondents identified as male, and 60 respondents identified as female. The respondents ranged in age from 18 to 68; the average age was 32.0, and the standard deviation was 8.5. On average, the respondents spent 13.42 min (5.95 min standard deviation) curating each countertop appliances list and 13.30 min (5.55 min standard deviation) curating each OTC medicine list.

4.5. Nonlinear optimization problem for smoke term curation

In automating the smoke term curation process, we first formally state the problem that we seek to solve. For each term *t* in the initial set of the smoke terms as identified by the CC scores [14], we define a

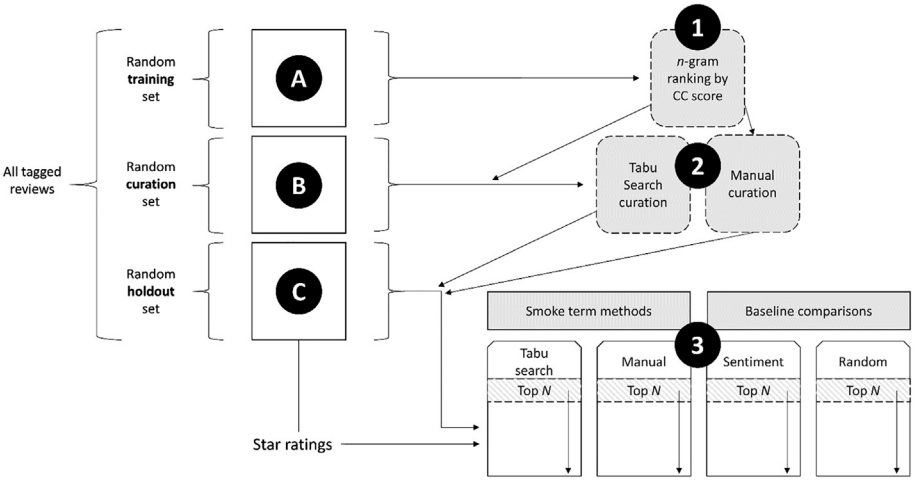

Fig. 1. Proposed data processing steps.

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

Table 1
Top-ranking *n*-grams from the training sample by CC score [14].

Rank	Unigram	CC score	Bigram	CC score	Trigram	CC score
Panel A: countertop appliances industry						
1	Off	155,069.14	Of the	139,112.67	All over the	104,577.72
2	Not	153,952.56	All over	119,251.24	Out of the	85,090.59
3	Dangerous	152,034.57	Be careful	110,028.00	Gets very hot	84,585.83
4	Started	146,861.34	On the	108,507.31	The first time	79,853.75
5	Fire	138,872.75	To the	106,779.68	Gets extremely hot	76,150.87
6	After	137,983.15	The bottom	103,991.15	Of the blade	75,457.09
7	Plastic	131,721.45	Out of	101,666.69	You have to	72,088.15
8	On	123,832.34	The plastic	100,026.68	Bottom of the	69,592.98
9	Hazard	123,262.19	It started	98,447.00	The bottom of	66,422.22
10	Out	122,144.02	The top	97,357.36	A fire hazard	66,054.64
Panel B: OTC medicine industry						
1	Studies	4966.67	Will I	6083.75	My stomach is	5212.88
2	RLS	4966.67	Stomach pain	5185.01	Not something I	5212.88
3	Stone	4966.67	Milk of	4966.67	To wear off	4966.67
4	Toll	4966.67	Leg syndrome	4966.67	Pain and cramping	4966.67
5	Capful	4966.67	Nasty I	4966.67	Afternoon I had	4966.67
6	Lack	4966.67	The women	4966.67	By afternoon I	4966.67
7	Magnesia	4966.67	Was fairly	4966.67	Never had this	4966.67
8	Enhance	4966.67	Of magnesia	4966.67	Just taking one	4966.67
9	Poisoning	4966.67	Restless leg	4966.67	Treatment for a	4966.67
10	Clog	4966.67	Strong so	4966.67	Maybe it was	4966.67

binary variable, x_t, that equals 1 if term t is included in the solution and 0 otherwise. We define S as a vector of the variables x_t for $t = 1…T$, or $[x_1\ x_2\ x_3…x_T]$. The user defines a value of T, representing the number of smoke terms to be considered. Suppose that the function $f(S,N)$ returns the number of defect-tagged reviews found in the top N-ranked reviews of the curation set using the smoke term list S and ranking the reviews using the CC scores as weights. Finally, the user defines a series of cutoffs for the top N-ranked reviews, n_i, where $i = 1…m$, and a series of associated weights for each of the cutoffs, w_i, where $\sum_{i=1}^{m} w_i = 1$. We define our nonlinear optimization problem as follows:

$$\text{Maximize} \sum_{i=1}^{m} w_i \left(\frac{f(S, n_i)}{n_i} \right)$$
$$x_t \text{ binary} \forall t$$

Note that this problem is nonlinear because the function $f(S,N)$ involves *ranking* the curation set of reviews using the smoke term list in S. For example, the smoke term list ["dangerous"] may yield 20 safety defects in the top 100-ranked reviews, and the smoke term list ["off"] may yield 10 safety defects in the top 100-ranked reviews. However, the objective values achieved by these lists are not linearly additive because they involve ranking; the smoke term list ["dangerous", "off"] may yield only 19 safety defects in the top 100-ranked reviews.

The user's ability to define a series of cutoffs and weights in this formulation has several benefits. First, from the perspective of a user, this allows additional flexibility by essentially allowing for multi-objective optimization of the number of defects observed within several arbitrary cutoffs. Second, these weights serve as tiebreakers for heuristic solvers. Suppose that $n_i = 100$, and $f(S_1,100) = f(S_2,100) = 30$. In this case, rather than forcing the algorithm to make a random choice between the two possibilities, we can further evaluate that $f(S_1,200) = 55 > f(S_2,200) = 50$. The former set of smoke terms, S_1, appears superior as, although it identifies the same number of safety defects in the top 100-ranked reviews, it identifies 5 more defects in the top 200-ranked reviews. Even if $n_i = 100$ is weighted as a more important cutoff, the solution in S_1 appears to push more safety defect-tagged reviews toward the top 100-ranked reviews, and the addition of some further term(s) in future iterations may shift these reviews into the top 100-ranked reviews. Maximizing performance at a series of cutoffs along a distribution systematically shifts safety defect-tagged reviews toward the top of the distribution.

4.6. Tabu search heuristic for smoke term curation

The Tabu search heuristic is a local search procedure that seeks to maximize an objective function by examining neighboring solutions until it reaches a local optimum at which all neighboring solutions result in inferior objective function values [37]. Rather than becoming "stuck" at local optima, the Tabu search algorithm allows movement in worsening directions if no improving moves are possible. A further attribute of the algorithm is *memory*: to ensure that the algorithm does not retest and cycle between the same potential solutions, previously explored solutions are disallowed (i.e., "Tabu"), to ensure that the algorithm explores additional feasible solutions.

Having delineated one-third of the reviews for each industry as curation sets, we implemented a Tabu search algorithm to provide high quality solutions to the aforementioned nonlinear program. In the following, we provide a segment of pseudocode for our Tabu search algorithm to choose those lists that maximize precision within the curation set. We equally weighted the precision in the top 50-ranked, top 100-ranked, and top 200-ranked reviews in our experiments [5,9–13], although we also obtained the same smoke term lists when experimenting with slightly different cutoffs and weighting schemes. Although these cutoffs are arbitrary, our observations of identical lists across different series of cutoffs suggest that optimal lists are rather resilient to choices of cutoffs.

In lines 1–5 of our pseudocode, we initialize a set of variables for the Tabu search algorithm. To best model effective responses to their specific situations, users may customize the values of *length*, *cutoffs*, and *weights*. In lines 8–10 of our pseudocode, we set each of the values of our initial smoke list, *solution*, equal to 0, indicating that each of the terms is excluded from the working smoke term list by default. This set of no smoke terms also serves as the initial value of *best_solution* in the algorithm. We run the body of the Tabu search algorithm until a stopping condition is reached. This condition is user-defined and may be a function of the amount of time that the algorithm has run, the number of iterations that have been run, or a lack of further improvement in the objective function. In our study, we observed promising results using the rule that the algorithm stopped if there had been no improvement in the previous 200 iterations. In lines 18–34, we test the effects of

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

changing the inclusion/exclusion status of each term upon the objective function. For every term excluded from the solution, we test the effect of adding that term to the solution; and for every term included in the solution, we test the effect of excluding that term from the solution. If this candidate solution is not in *tabu_list*, then we evaluate its fitness and add it to our running record of potential term lists to pivot to in the next iteration. After evaluating each of the potential one-step moves from the current solution, we pivot from the current solution to the highest performing candidate solution, and we add the highest-performing candidate solution to *tabu_list* to ensure that the algorithm does not cyclically revert to this solution. If the objective value achieved by the highest performing candidate solution outperforms the objective value achieved by the highest performing solution previously found, then this current solution is recorded in *best_solution*.

The principle employed in lines 40–42 is an example of a greedy heuristic [37], a particular variety of local search procedure that chooses the decision variable that provides the greatest improvement. Conceptually, it is also similar to a best-first search for graph exploration [38]. Although potentially many of the possible changes to the working smoke term list may result in an improvement, the algorithm chooses the most locally satisfying option. This heuristic cannot guarantee a global optimum solution, but it does improve the objective function in large leaps, and the Tabu search rules help to prevent the algorithm from becoming satisfied with local optima.

Although testing multi-step moves rather than only one-step moves would alleviate some of the limitations of the greedy heuristic, doing so would greatly add to the computational requirements of the problem. Consider that testing all one-step moves requires *at most* 200 tests

```
1   declare integer length ← 200 // Number of smoke terms to consider
2   declare array cutoffs ← [50, 100, 200] // Cutoffs considered (n(i))
3   declare array weights ← [0.333, 0.333, 0.333] // Weights considered (w(i))
4   declare array solution ← [ ]
5   declare array tabu_list ← [ ]
6
7   // Define initially empty smoke term list
8   for i = 1 to length
9       solution.add(0)
10  end for
11  declare array best_solution ← solution
12
13  while not (stopping_condition()) // Time elapsed without improvement in objective
14
15      declare array candidates ← [ ]
16      declare array fitnesses ← [ ]
17
18      for i = 1 to length
19
20          // Generate adjacent solution
21          declare array candidate ← solution
22          if candidate[i] = 1 then
23              candidate[i] ← 0
24          Else
25              candidate[i] ← 1
26          end if
27
28          // Evaluate fitness of smoke term list
29          if not (candidate in tabu_list) then
30              fitnesses.add(fitness(candidate, cutoffs, weights))
31              candidates.add(candidate)
32          end if
33
34      end for
35
36      declare integer index = argmax(fitnesses) // Determine best candidate
37      solution ← candidates[index]
38      tabu_list.add(solution)
39
40      if fitnesses[index] > fitness(best_solution, cutoffs, weights) then
41          best_solution ← solution // Update best solution
42      end if
43
44  end while
45
46  return best_solution
```

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

because the set of possible smoke term choices has 200 possible entries. In a two-step test, 200 additional tests must be performed within each of the 200 initial tests, resulting in an additional 40,000 tests, or 40,200 total tests. Many of these permutations are identical combinations, so only a maximum of 20,100 tests must be evaluated once duplicates are removed, or 101.5 times as many tests as in the one-step case. If we to expand to three-step moves, then a maximum of 1,333,500 tests are required for a single iteration.

Although for brevity we do not report the results in this paper, we also attempted the use of a genetic algorithm [39] to perform the same task of smoke term curation. We found the performance of this algorithm to be inferior to the results achieved with the Tabu search algorithm. As genetic algorithms are based on retaining random changes in decision variables that improve the solution, we found that our algorithm often failed to capitalize upon narrow but extremely profitable paths that the greedy heuristic in the Tabu search algorithm easily identified.

4.7. Incorporating non-textual data

Much of the literature has made use of text analytics such as sentiment analysis [26,27] and word frequency [5,9–13] in order to derive insights from textual data. In many cases, online media mainly only provide data in a textual format, making text mining techniques the only appropriate tools for analysis. However, online reviews are unique in that they frequently include additional data, such as a rating of the product in "stars", images of the product being reviewed, or other users' ratings of the helpfulness of the review. Star ratings are the most ubiquitous of these data types in online reviews, providing a numerical summary of a consumer's experience with a product. Importantly, star ratings may provide different information than the textual content of reviews. Mudambi et al. [19] study misalignment between the sentiment expressed in the text bodies of online reviews and star ratings. The authors argue that misalignment is particularly prevalent in reviews that contain high star ratings because users feel the need to balance out positive content with negative content to maintain credibility [19], a phenomenon also acknowledged by the psychology literature [40]. As star ratings are often predictive of product quality [7], incorporating them in addition to textual content may improve the performance of our technique.

Recognizing the potential of additional online review characteristics, specifically star ratings, to add to the abilities of text mining techniques to decipher the presence of safety defects, we seek to integrate this data into our methodology. First, as our star ratings from Amazon.com are scaled from 1 to 5, a rating of 1 indicates the worst-perceived quality of products, while a rating of 5 indicates the best-perceived quality of products. We create a new measure, *inverted star rating*, or (6 – star rating), so that high values of the measure instead reflect the worst quality of products. We perform this step to maintain consistency with our CC score metric, on which high values reflect greater prevalence of safety defect-related language. Recall that the scores serve as weights upon each term in the final analysis such that a review's total score is incremented by the corresponding CC score each time that a relevant term occurs within its text. To obtain what we refer to as an augmented score, we multiply each review's total score by its inverted star rating. Using this formulation models the relationship between star ratings and textual content as an interaction effect such that reviews that score high values on both metrics score especially highly in their augmented scores. This step should counter situations in which reviews with high star ratings also make negative comments, as these reviews will have low inverted star scores; meanwhile, augmented scores for reviews with safety-defect related text and low star ratings (high inverted star ratings) are accentuated, causing these reviews to be ranked very highly.

5. Results and evaluation

In Table 2, we display the 10 most commonly chosen unigrams, bigrams, and trigrams from our samples of manual curators. In general, the terms chosen most frequently by these manual curators seemed consistent with methodologies employed in prior work [11–13]. However, particularly in the OTC medicine industry, we observed that curators made considerably different choices as to which terms to retain. Participants were instructed to choose a list of terms that they thought best delineated safety defects, but they were not otherwise instructed on a process to use for smoke term curation. Despite this, the most frequently chosen terms typically reflect removal of common English words or "stop words"; product and brand names were not chosen especially frequently; and sub-product terms were not chosen especially frequently. Beyond these principles, each curator further subjectively assessed which terms they believed to be relevant. Many of the terms seem to relate clearly to a type of safety hazard that a product may cause, such as "fire hazard" for countertop appliances or "stomach pain" for OTC medicine. Some terms, such as "day I was" interestingly seem to invoke a narrative, which is common in reviews referring to safety defects. For example, one review in our countertop appliances sample states the following.

"*On the four* ***[day I was]*** *mixing, not an overly long length of time, when I smelled that smell you get when something electrical is burning.*"

We evaluated the performance of each of the smoke term lists generated by the manual curators in our holdout sample. Furthermore, we evaluated the performance of those smoke term lists again while incorporating the star rating augmentation in our formula. We display the results of these evaluations in Table 3 and Table 4. Predictably, performance was superior for the countertop appliances industry as the much larger sample size increased the number of defects possible for the algorithms to find. Across the three lengths of *n*-grams, results were fairly consistent. On average, curators supplied smoke term lists containing 30–32 terms for countertop appliances or 17–19 terms for OTC medicine. Before score augmentation, these lists generally found about 16/32/60 defects in the top 50/100/200-ranked reviews for countertop appliances and about 6/7/11 for OTC medicine. Importantly, we also observed substantial variability in the performance of smoke term lists, indicating that the choice of smoke terms greatly affects the final results in this methodology. Although we found that the standard deviation of the number of safety defects found in the top *N*-ranked reviews increased as the cutoff increased, this appeared to reflect larger maxima in the expanded dataset as each cutoff threshold was relaxed (from 50 to 100 to 200) rather than a flatter distribution. We did not observe any meaningful difference in performance based on the demographic data collected, including gender identity, age, and educational background.

After augmenting scores using star ratings, we observed a substantial improvement for each type of list, although this improvement seemed to be strongest when considering shorter *n*-grams. Indeed, we verified that precision for each of the manually curated lists improved or stayed the same at each of the chosen cutoffs. To assess this improvement, we performed pairwise X^2 tests to compare the proportion of safety defect-tagged reviews that each list detected at each threshold to the proportion detected after score augmentation. We display the results of these tests in Table 5. Many of the unigram lists significantly differed at the 0.05 level after incorporating star ratings, but the proportion of lists that significantly differed declined as the lengths of the *n*-grams increased.

In Tables 6–8, we provide the lists of unigrams, bigrams, and trigrams curated by the Tabu search algorithm. For each term, we provide the CC score, that CC score's rank relative to the CC scores of the other *n*-grams in the initial list, and a rank of how frequently the term appeared in the lists generated by manual curators. We observed only moderate

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

Table 2
Most commonly curated *n*-gram smoke terms by manual curators.

Rank	Unigram	Count	Bigram	Count	Trigram	Count
Panel A: countertop appliances industry ($N = 102$)						
1	Dangerous	73 (72%)	Fire hazard	73 (72%)	Top of the	73 (72%)
2	Hazard	72 (71%)	Extremely hot	61 (60%)	Will never buy	61 (60%)
3	Burned	70 (69%)	Burning smell	60 (59%)	I noticed the	60 (59%)
4	Fire	69 (68%)	Is dangerous	58 (57%)	Unplugged it and	58 (57%)
5	Smoke	65 (64%)	Caught fire	57 (56%)	Is a cheap	57 (56%)
6	Broke	65 (64%)	Very hot	57 (56%)	Not recommend this	57 (56%)
7	Crack	62 (61%)	Burned my	54 (53%)	Day I was	54 (53%)
8	Leaking	62 (61%)	Get burned	54 (53%)	It gets very	54 (53%)
9	Burning	60 (59%)	Smoke was	53 (52%)	Waste of money	53 (52%)
10	Melted	59 (58%)	Burning yourself	51 (50%)	I had the	51 (50%)
Panel B: OTC medicine industry ($N = 139$)						
1	Poisoning	90 (65%)	Stomach pain	87 (63%)	Pain and cramping	84 (60%)
2	Damage	80 (58%)	Health concerns	66 (47%)	Would not recommend	54 (39%)
3	Unsafe	60 (43%)	Food poisoning	57 (41%)	Restless leg syndrome	47 (34%)
4	Exceed	56 (40%)	Not recommend	54 (39%)	Burns your throat	45 (32%)
5	Caution	51 (37%)	Caution if	53 (38%)	Multi symptom allergy	44 (32%)
6	Overdosed	48 (35%)	With caution	48 (35%)	Not recommend this	42 (30%)
7	Faint	43 (31%)	Cold sweats	48 (35%)	Exceed the recommended	42 (30%)
8	Sweats	42 (30%)	Cause drowsiness	45 (32%)	Having trouble sleeping	39 (28%)
9	Liver	41 (29%)	Symptom allergy	42 (30%)	Abdominal pain mild	37 (27%)
10	Potentially	38 (27%)	Have caused	42 (30%)	With caution if	37 (27%)

agreement between the CC score rankings and the manual curation frequency rankings of these selected smoke terms. One striking aspect of these curation decisions is that the lists curated by Tabu search are shorter than the typical human-curated lists. While each human-curated list averaged around 30–32 terms for countertop appliances or 17–19 terms for OTC medicine, our Tabu search recommended unigram, bigram, and trigram lists contained 11, 21, and 26 terms for countertop appliances and 11, 9, and 12 terms for OTC medicine. It appears that the algorithm recommended more targeted smoke term lists than human curators; rather than attempting to include every possibly applicable smoke term, our algorithm only chose the smoke terms that had marginal effectiveness relative to the incumbent solution. For example, it is possible that some smoke terms are frequently used in combination, so adding both terms to a final solution may not improve precision very much, or precision may worsen if one of the terms has an alternative sense in reviews not referring to safety defects. As bigram and trigram phrases represent quite specific strings of text in the human language, our algorithm generally chose greater numbers of these terms to build effective lists, whereas unigrams may be more ubiquitous. This aspect is consistent with Zipf's law [41], which posits that a term's frequency is inversely proportional with its frequency table rank. Therefore, we might expect a small set of words to occur quite frequently in reviews indicating safety defects.

Table 3
Performance of manually curated *n*-gram smoke term lists: countertop appliances industry.

		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50[a]	Top 100[a]	Top 200[a]
Panel A: performance of manually curated unigrams ($N = 102$).							
Minimum	1.00	8.00	17.00	29.00	14.00	25.00	54.00
Median	28.00	14.50	31.00	63.00	25.00	50.00	97.50
Mean	30.27	14.95	30.70	60.36	24.67	48.64	94.84
Maximum	115.00	32.00	49.00	92.00	36.00	67.00	125.00
Standard deviation	20.91	4.13	7.48	14.74	5.02	9.45	18.03
Panel B: performance of manually curated bigrams ($N = 102$).							
Minimum	1.00	7.00	14.00	32.00	11.00	22.00	32.00
Median	25.50	16.00	30.50	61.50	22.00	41.50	83.50
Mean	31.74	16.37	31.84	60.63	22.75	43.36	83.68
Maximum	140.00	33.00	52.00	92.00	36.00	65.00	117.00
Standard deviation	26.58	5.14	8.65	15.10	5.85	9.93	15.76
Panel C: performance of manually curated trigrams ($N = 102$).							
Minimum	1.00	5.00	11.00	20.00	6.00	11.00	20.00
Median	23.00	17.00	31.00	61.00	21.00	39.50	74.00
Mean	30.38	16.64	31.81	58.65	20.61	38.64	71.32
Maximum	175.00	32.00	46.00	87.00	32.00	62.00	95.00
Standard deviation	30.46	4.69	7.87	14.81	4.51	8.12	15.43

[a] Indicates that scores were augmented by star ratings.

Table 4
Performance of manually curated *n*-gram smoke term lists: OTC medicine industry.

		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50[a]	Top 100[a]	Top 200[a]
Panel A: performance of manually curated unigrams ($N = 139$).							
Minimum	1.00	1.00	2.00	5.00	3.00	5.00	7.00
Median	12.00	6.00	7.00	11.00	7.00	8.00	12.00
Mean	16.96	6.01	7.42	11.01	7.50	8.90	12.62
Maximum	100.00	13.00	15.00	20.00	15.00	18.00	31.00
Standard deviation	17.32	3.12	3.57	3.67	3.75	4.66	5.45
Panel B: performance of manually curated bigrams ($N = 139$).							
Minimum	2.00	1.00	2.00	5.00	3.00	5.00	7.00
Median	14.00	6.00	7.00	10.00	7.00	8.00	12.00
Mean	18.96	5.95	7.11	10.41	7.35	8.52	11.96
Maximum	128.00	12.00	14.00	18.00	14.00	18.00	29.00
Standard deviation	19.48	3.22	3.45	3.59	3.71	4.50	5.20
Panel C: performance of manually curated trigrams ($N = 139$).							
Minimum	1.00	1.00	2.00	5.00	3.00	5.00	7.00
Median	11.00	6.00	7.00	11.00	7.00	8.00	12.00
Mean	17.71	5.81	6.75	10.61	7.25	8.22	12.17
Maximum	129.00	11.00	13.00	18.00	14.00	17.00	29.00
Standard deviation	20.40	3.00	3.41	3.63	3.84	4.57	5.33

[a] Indicates that scores were augmented by star ratings.

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

Table 5
Pairwise comparisons of manually curated *n*-gram lists after score augmentation.

n-Gram list type	Cutoff	Count of lists differing by X^2 test at the 0.05 level	
		Countertop appliances industry	OTC medicine industry
Unigrams	50	85 (83%)	23 (17%)
	100	86 (84%)	27 (19%)
	200	96 (94%)	32 (23%)
Bigrams	50	29 (28%)	24 (17%)
	100	46 (45%)	27 (19%)
	200	70 (69%)	30 (22%)
Trigrams	50	10 (10%)	15 (11%)
	100	31 (30%)	18 (13%)
	200	36 (35%)	21 (15%)

Another interesting aspect of these curated lists concerns the CC score ranks of each of the selected terms. While the algorithm certainly makes use of some terms with the highest ranks within the top 200 CC scores, we observed that the algorithm actually incorporated a range of terms that spanned the selection of the top 200 CC scoring terms fairly completely. Interestingly, when we attempted expanding the set of allowable smoke terms beyond 200 terms, we observed a decline in quality of terms consistent with prior research [10,11], and the algorithm did not select these additional terms for inclusion. This aspect of the result illustrates the value of the CC score in generating smoke term lists; without using a technique like the CC score, the process of narrowing down the potential *n*-grams to potential candidates would be enormous. However, limiting the curation process to 200 candidate terms substantially reduces the computational requirements of the problem.

The ranks of the smoke terms in the manual curation experiment reveal further insight as to the differences between human-curated and machine-curated smoke term lists. Whereas both human-curated and machine-curated lists frequently tended to focus on aspects of products that may be hazardous (e.g., "dangerous", "fire hazard", "pain and cramping"), the machine-curated smoke term lists also included some terms that seemed to refer to customer narratives. In reviews describing experiences with safety defects, customers frequently offer narratives of their experiences with products that led to dissatisfaction. The machine-curated smoke lists capture these experiences with terms like "touched" and "that evening" that customers typically do not use unless they are explaining experience with a safety defect. However, these terms may be far less obvious from the perspective of human curators. For example, the weak auxiliary term "began" in the machine-curated countertop appliances unigram list was rarely chosen in the human-curated smoke term lists relative to stronger nouns, verbs, and adjectives, but it is frequently used as customers explain their experiences. The following example review uses "began" to establish the narrative of the customer's experience with a product.

> "I've had my (product name) set for quite a while and liked it very much HOWEVER today as I [**began**] to heat it, it exploded. The two parts separated and flew in opposite direction. I was hit in the head, the burners were pushed apart and the lights over the stove were wrecked."

We were encouraged to find that the performances of each of these smoke term lists on the holdout sample were excellent. Each smoke term list outperforms both the average comparable human-curated list and the highest-performing comparable human-curated list at each of the three cutoffs considered. The Tabu search algorithm seems to generalize across industries, and it improved performance for the large sample in the countertop appliances industry and the smaller sample of the OTC medicine industry. Like the human-curated smoke term lists, we find that augmenting the scores with star ratings improves the performance of each smoke term list, but the improvement is greatest for the unigram list. We detail the performance of these smoke term lists in Table 9. Interestingly, we observed that the performances of smoke term lists are extremely sensitive to small changes. Consider if the word "hot" for countertop appliances, which was ranked 18th by the CC score metric and was commonly included in human-

Table 6
List of unigrams curated by the Tabu search algorithm.

Unigram	CC score	CC score rank	Manual curation rank
Panel A: countertop appliances industry			
Dangerous	152,034.57	3	1
Fire	138,872.75	5	4
Hazard	123,262.19	9	2
Recalled	91,334.24	49	25
Safety	86,857.95	55	32
Caught	75,736.65	92	63
Began	69,614.20	127	183
Touched	68,696.00	134	108
Cracked	66,735.64	149	15
Defect	64,757.91	166	12
Leak	62,440.46	181	28
Panel B: OTC medicine industry			
Capful	4966.67	5	45
Caution	4626.46	13	5
Liver	4507.69	17	9
Potentially	3990.57	27	10
Surrounding	3990.57	28	160
Caused	3567.76	32	11
Catastrophic	3511.48	105	14
Overdosed	3511.48	109	6
Chills	3511.48	120	47
Monster	3511.48	195	104
Heaving	3511.48	200	38

Table 7
List of bigrams curated by the Tabu search algorithm.

Bigram	CC score	CC score rank	Manual curation rank
Panel A: countertop appliances industry			
Fire hazard	95,425.81	11	1
On fire	93,418.02	14	35
Extremely hot	89,266.37	22	2
Burned my	80,784.65	34	7
A fire	79,092.70	38	30
Gets extremely	76,150.87	47	29
And smoke	75,421.36	50	17
Too hot	73,937.67	55	19
Gets hot	71,931.19	60	24
Smoke was	71,339.42	61	9
Burning smell	69,037.87	67	3
Caught fire	67,785.14	71	5
Is dangerous	66,442.69	79	4
Design flaw	65,001.78	89	14
Dangerous I	61,863.60	106	15
You're stuck	56,244.37	157	70
Shattered into	56,244.37	158	44
Smoke started	56,244.37	160	11
Burning yourself	55,181.12	173	10
Your fingers	54,693.41	182	68
Started leaking	54,422.37	187	18
Panel B: OTC medicine industry			
Stomach pain	5185.01	2	1
Caution if	4966.67	15	5
By afternoon	4966.67	18	124
It caused	4966.67	19	12
With caution	4612.61	48	6
Periods of	3990.57	79	67
That evening	3990.57	90	96
Severely burnt	3511.48	147	24
Your belly	3511.48	166	146

Table 8
List of trigrams curated by the Tabu search algorithm.

Trigram	CC score	CC score rank	Manual curation rank
Panel A: countertop appliances industry			
Gets very hot	84,585.83	3	17
Gets extremely hot	76,150.87	5	134
A fire hazard	66,054.64	10	80
Got so hot	61,542.29	21	82
The handle gets	57,114.19	32	27
House on fire	56,244.37	40	62
It started leaking	52,846.59	55	15
So hot that	52,060.14	60	23
Hand on the	51,648.10	65	197
After 3 uses	50,439.18	72	44
Caught on fire	50,018.20	78	76
Unplugged it and	50,018.20	79	4
Broke after about	47,244.47	107	22
Smell and taste	47,101.96	110	29
All over my	46,936.64	111	104
Is no way	46,277.70	114	90
Catch on fire	45,667.22	122	175
Hot on the	45,280.00	135	170
Gets really hot	45,062.89	140	151
The lid broke	44,498.13	149	49
Then it started	44,126.46	163	154
To leak after	43,957.87	165	182
I burned my	43,957.87	173	11
Handle gets hot	43,957.87	176	140
Started leaking water	43,957.87	182	68
Burned my fingers	43,957.87	183	74
Panel B: OTC medicine industry			
My stomach is	5212.88	1	17
Pain and cramping	4966.67	4	1
Wouldn't use it	4966.67	16	16
Going to faint	4966.67	19	17
Had this problem	4966.67	31	21
And very bad	4966.67	37	23
With caution if	4966.67	38	10
It with caution	4966.67	42	11
This product contains	3990.57	57	25
Sweats and flu	3511.48	90	15
All day now	3511.48	158	143
Morning and still	3511.48	171	135

curated smoke term lists (41/102 unigram lists), were added to the machine-curated unigram list. In such a case, we would observe a drop in precision by more than 40% at each cutoff, falling to 12/24/62 safety defect-tagged reviews in the top 50/100/200-ranked reviews. Although "hot" was used frequently in safety defect-tagged reviews as a component of the machine-curated bigram and trigram lists (e.g., "gets hot", "extremely hot", "gets extremely hot"), the unigram alone introduces many false positives. Without context provided by the accompanying words in the bigrams and trigrams, "hot" may actually reflect a positive quality. This particular nuance is vital to the performance of smoke term lists, but it is nearly impossible for manual curators to predict, verifying the value of machine-curated curation. Consider the following example review containing the term "hot".

> "Does everything it says it'll do.No problems with this pot.It keeps the coffee [**hot**] at home & on the go."

Our results seem to confirm the expectations noted by prior researchers that smoke terms are highly industry-specific [5,9–13]. Broadly, we observed three types of smoke terms: industry-specific hazard indicators, such as "fire hazard" or "heaving"; general hazard indicators, such as "dangerous" and "with caution"; and industry-specific narrative terms, such as "hand on the" or "by afternoon". For example, "hand on the" might be a part of the narrative, "I burned my [**hand on the**] toaster" in a countertop appliance review, and "by afternoon" might be a part of the narrative, "I was vomiting [**by afternoon**]" in an OTC medicine narrative. Although neither term directly references a safety defect, each tends to be used in the context of a safety-related narrative. Interestingly, even though terms such as "dangerous" and "with caution" seem applicable across industries, customers' uses of these general terms appeared to be quite industry-specific, as there were no overlaps between the two sets of industry-specific smoke terms lists that we generated. We observed that each industry-specific smoke term list performed quite poorly when applied to the other industry, as its terms were not very predictive in alternate domains. We also attempted an experiment in which we mixed countertop appliance and OTC medicine reviews together before applying our algorithm, but the resulting smoke term lists performed poorly both on the combined sample and on the industry-specific samples as each industry introduced noise relative to the other. As such, our experiments support the findings of prior research that industry-specific smoke term lists are necessary for safety surveillance.

To provide further statistical validation for our algorithm's improvement upon human-curated smoke terms, we further performed a series of X^2 tests comparing the proportions of defects found at each cutoff between the machine-curated smoke term lists and the comparable human-curated smoke term lists. We found statistical evidence that the proportions obtained by our algorithm differed from the proportions obtained by the average human-curated smoke term lists at the 0.05 level for each *n*-gram type and at each cutoff across both industries.

Finally, we sought to assess whether the score augmentation resulted in a statistically significant difference between comparable machine-curated smoke term lists. For countertop appliances, we found that the pre-augmentation unigrams list performance differed at the 50, 100, and 200 cutoffs from the post-augmentation unigrams list performance. The pre-augmentation bigrams list performance differed at the 100 cutoff from the post-augmentation bigrams list performance. Finally, the pre-augmentation trigrams performance did not significantly differ from the post-augmentation trigrams performance at any of the cutoffs tested. We observed a significant difference post-augmentation for the OTC medicine industry for the trigrams list at the 200 cutoff.

In Figs. 2–4, we provide series of ROC curves showing the performance of each machine-curated smoke term list at a range of cutoff values for the top *N*-ranked reviews. We focus our charts on the top-ranking parts of the distribution. In addition to showing machine-curated smoke term lists, we also show AFINN and Harvard GI sentiment analyses [22,24] and a random chance baseline. The degree of "lift" in each chart shows that machine-curated smoke term lists augmented by star ratings offer the best performance.

Table 9
Performance of machine-curated *n*-gram smoke term lists.

		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50[a]	Top 100[a]	Top 200[a]
Panel A: countertop appliances industry							
Unigrams	11	36	63	108	44	76	130
Bigrams	21	35	60	102	39	70	129
Trigrams	26	34	58	102	35	65	115
Panel B: OTC medicine industry							
Unigrams	11	15	21	32	20	28	43
Bigrams	9	14	21	30	15	24	38
Trigrams	12	14	20	29	17	25	42

[a] Indicates that scores were augmented by star ratings.

6. Limitations

The process of tagging reviews used in this paper required an ample level of human effort, and tagging methods have become standard in

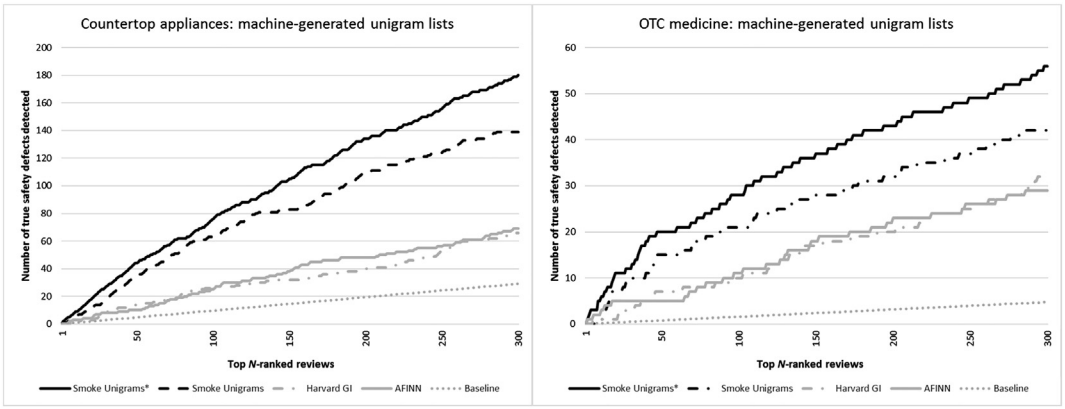

Fig. 2. Performance of machine-curated unigram lists.

defect discovery literature for the development of training data [5,9–13]. However, we acknowledge that the process of tagging reviews is not without subjectivity. We observed considerable agreement amongst our taggers, but we recognize that these designations are still not immune to training biases. As such, we recommend that practitioners supplement our techniques with further data to ensure that defect discovery is comprehensive. Additionally, the tagging procedure used in this paper employed review-level specificity in the sense that entire reviews were tagged as either indicating safety defects or not indicating safety defects. A more granular technique might involve tagging shorter sub-sections of reviews, such as sentences or phrases. This technique may further ameliorate false positives by ensuring that surrounding terms that do not reflect safety defects are not flagged by the CC scoring metric due to high prevalence in safety defect-tagged reviews. Given the enormous volume of online review data tagged in this study, implementing this technique may necessitate an onerous labor requirement.

The Tabu search algorithm implemented in this paper is admittedly a heuristic that cannot guarantee globally optimal results to problems without exhaustive enumeration. Relative to a greedy heuristic on its own, however, the Tabu search algorithm allows the solver to explore more of the feasible region after reaching local optima. Although we cannot guarantee globally optimal solutions, we are satisfied with our algorithm's performance given that it clearly provided improved precision relative to the status quo of human-curated smoke term lists.

Finally, we note that the defect discovery technique discussed in this paper should constitute a profoundly useful tool for manufacturers and regulatory agencies, but it cannot entirely supplant existing methods of detection. The population of users that post reviews or other product feedback online is large and growing, but it still represents a subsample of the entire population of consumers [8]. Additional offline screening methods, such as warranty claim analysis, consumer surveys, safety complaint filings with regulators, and physical testing of products are vital defect prevention methods to be employed in addition to the discussed techniques.

7. Conclusion and implications

In this study, we develop a novel methodology for smoke term curation by replacing a formerly manual and subjective term-list curation task with a Tabu search algorithm. Using a sample of human-curated smoke term lists, we found that human curation is highly variable, motivating the need for a more objective automated system. Given the same dataset, the Tabu search algorithm provided great improvements upon each comparable human-curated list and more objective curation based on precision in the curation set. We found that incorporating non-textual star ratings in our analysis improved precision of human-curated and machine-curated smoke term lists.

Our study has wide-ranging implications for industry, regulatory agencies, and researchers. One major implication of our study is that

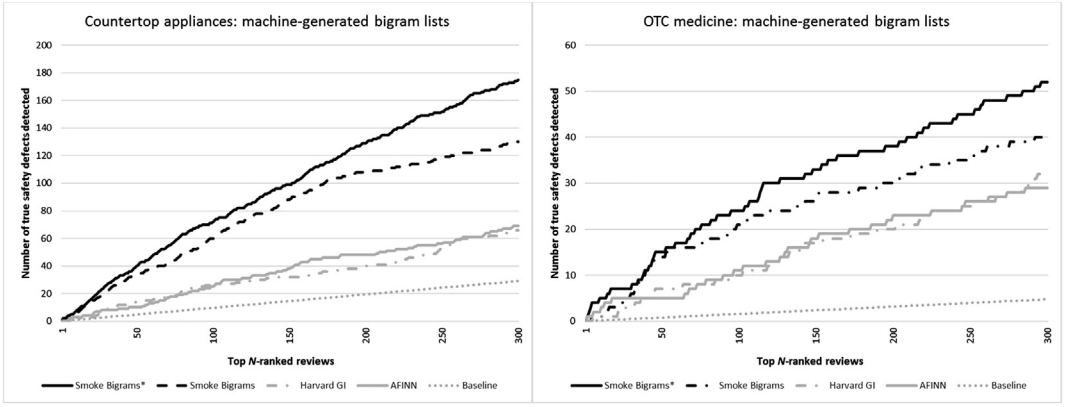

Fig. 3. Performance of machine-curated bigram lists.

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

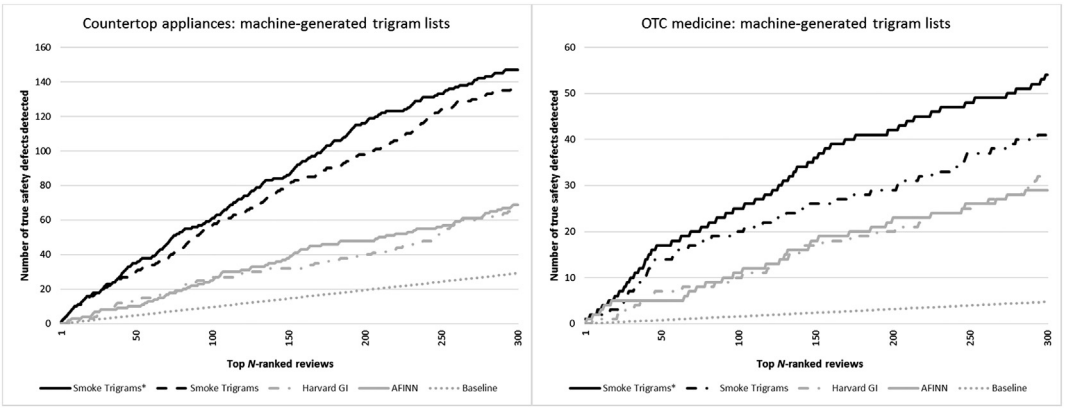

Fig. 4. Performance of machine-curated trigram lists.

heuristics like the Tabu search algorithm prove effective tools for improving the performance of defect discovery techniques. The new possibility for organizations to make use of these heuristics in place of manual curation should allow superior performance and more effective and rapid detection of safety defects. We observed that the performance of smoke term lists is sensitive to changes in the list, so manual curation may not offer stable performance. We recommend that researchers and practitioners use large datasets of online reviews, as these will span the most unique terms and phrases and allow the algorithm to more thoroughly capture the nuances of human language. With a small sample size, some of these subtle effects may be difficult to detect. Additionally, relative to machine-curated lists, human-curated lists were far less likely to include terms such as "began" or "hand on the" that reflect elements of a consumer's narrative describing their experience with a product. These narratives may be a signal of the discussion of safety defects in online reviews; further study may attempt to understand the extent to which this narrative structure provides additional insights for defect discovery. Furthermore, the use of an algorithm to perform smoke term curation removes a potentially arduous labor requirement for a human to sift through potential smoke term candidates and choose what they believe to be the best possible smoke term list. The performance of human-curated smoke term lists is highly variable, and the use of a heuristic algorithm offers a more stable and reliable level of performance for this vital task.

A further implication of our study reflects the potential value of non-textual data in defect discovery. Most contemporary works in this area have focused exclusively on the textual characteristics of reviews [5,9–13]; however, we found that including star ratings to augment scores obtained by smoke terms resulted in statistically significant improvements in performance. We hope that this finding provides impetus for further research on creative ways to integrate this data with existing techniques in search of further improvements in performance.

A final implication of our study specifically affects defect discovery in the countertop appliances and OTC medicine industries. The smoke term lists provided in this paper are actionable guidelines for firms for detecting defects in online reviews. As earlier studies have found [5,9–13], smoke term lists appear domain-specific, including terms such as "cracked" or "heaving" that may not apply in other industries. These lists provide insights as to types of safety defects experienced most often and may be used for discovering safety defects and developing responses.

Acknowledgements

The authors are grateful to Rich Gruss, lead developer of the PamTag collaborative tagging system, for providing access to PamTag. The authors are also grateful to Siriporn Srisawas for her assistance managing the team of student taggers for the over-the-counter (OTC) medicine dataset.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.dss.2017.10.012.

References

[1] Y. Lee, Samsung Note 7 Recall to Cost at Least $5.3 Billion, Associated Press, October 14, 2016.
[2] Y. Hagiwara, T. Taniguchi, Takata Puts Worst-case Airbag Recall Costs at $24 Billion, Bloomberg, March 30, 2016.
[3] N.G. Rupp, The attributes of a costly recall: evidence from the automotive industry, Rev. Ind. Organ. 25 (2004) 21–44.
[4] United States Consumer Product Safety Commission (CPSC), 2016 Annual Report, https://www.cpsc.gov/content/2016-annual-report June 9, 2017.
[5] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decis. Support. Syst. 54 (2012) 87–97.
[6] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: online book reviews, J. Mark. Res. 43 (2006) 345–354.
[7] N. Hu, L. Liu, J.J. Zhang, Do online reviews affect product sales? The role of reviewer characteristics and temporal effects, Inf. Technol. Manag. 9 (2008) 201–214.
[8] BrightLocal, Local Consumer Review Survey, https://www.brightlocal.com/learn/local-consumer-review-survey/ 2016, Accessed date: 23 August 2017.
[9] A.S. Abrahams, W. Fan, G.A. Wang, Z.J. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery, Prod. Oper. Manag. 24 (2015) 975–990.
[10] A.S. Abrahams, J. Jiao, W. Fan, G.A. Wang, Z. Zhang, What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings, Decis. Support. Syst. 55 (2013) 871–882.
[11] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decis. Support. Syst. 90 (2016) 23–32.
[12] D. Law, R. Gruss, A.S. Abrahams, Automated defect discovery for dishwasher appliances from online consumer reviews, Expert Syst. Appl. 67 (2017) 84–94.
[13] D.Z. Adams, R. Gruss, A.S. Abrahams, Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews, Int. J. Med. Inform. 100 (2017) 108–120.
[14] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decis. Support. Syst. 40 (2005) 213–233.
[15] ConsumerReports, Appliance fires pose a safety concern, https://www.consumerreports.org/cro/magazine/2012/03/appliance-fires-is-your-home-safe/index.htm March 2012, Accessed date: 23 August 2017.
[16] B. Perlow, 22 Models of Cuisinart food processors recalled after reports of blade breaking off, ABC News, December 13, 2016.
[17] D.J. Neal, 3 Children's medicines recalled because of potentially lethal ingredient, Sacramento Bee, November 25, 2016.
[18] W. Duan, B. Gu, A.B. Whinston, The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry, J. Retail. 84 (2008) 233–242.
[19] S.M. Mudambi, D. Schuff, Z. Zhang, Why aren't the stars aligned? An analysis of online review content and star ratings, 47th Hawaii International Conference on System Sciences, IEEE 2014, pp. 3139–3147.
[20] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: a sentiment analysis approach, Decis. Support. Syst. 55 (2013) 919–926.
[21] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2011) 1–8.
[22] F.Å. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, Proceedings of the 1st Workshop on Making Sense of Microposts 2011, pp. 93–98.
[23] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEW): instruction Manual and Affective Ratings, Citeseer, 1999.
[24] E.F. Kelly, P.J. Stone, Computer Recognition of English Word Senses, North-Holland, 1975.

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012

[25] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Am. Soc. Inf. Sci. Technol. 63 (2012) 163–173.
[26] S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, J. Manag. Inf. Syst. 29 (2013) 217–248.
[27] H. Tang, S. Tan, X. Cheng, A survey on sentiment detection of reviews, Expert Syst. Appl. 36 (2009) 10760–10773.
[28] H. Isah, P. Trundle, D. Neagu, Social media analysis for product safety using text mining and sentiment analysis, 14th UK Workshop on Computational Intelligence, IEEE 2014, pp. 1–7.
[29] C.C. Yang, H. Yang, L. Jiang, M. Zhang, Social media mining for drug safety signal detection, Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, ACM 2012, pp. 33–40.
[30] S.E. Robertson, On relevance weight estimation and query expansion, J. Doc. 42 (1986) 182–188.
[31] G. Salton, The SMART Retrieval System—Experiments in Automatic Document Processing, Prentice Hall, 1971.
[32] H.T. Ng, W.B. Goh, K.L. Low, Feature selection, perceptron learning, and a usability case study for text categorization, ACM SIGIR Forum, ACM 1997, pp. 67–73.
[33] J. McAuley, R. Pandey, J. Leskovec, Inferring networks of substitutable and complementary products, Proceedings of the 21th SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM 2015, pp. 785–794.
[34] J. Cohen, Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit, Psychol. Bull. 70 (1968) 213.
[35] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[36] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, John Wiley & Sons, 2013.
[37] F. Glover, Future paths for integer programming and links to artificial intelligence, Comput. Oper. Res. 13 (1986) 533–549.
[38] R. Dechter, J. Pearl, Generalized best-first search strategies and the optimality of A*, J. ACM 32 (1985) 505–536.
[39] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.
[40] T.M. Amabile, Brilliant but cruel: perceptions of negative evaluators, J. Exp. Soc. Psychol. 19 (1983) 146–156.
[41] G.K. Zipf, The Psycho-biology of Language, Houghton Mifflin, 1935.

David M. Goldberg is a Ph.D. student in the Department of Business Information Technology at the Pamplin College of Business at Virginia Tech. He received a B.S. in Business Information Technology and a B.A. in Geography at Virginia Tech. He has previously worked in the commercial real estate industry, where he developed spatial tools to augment market analyses. His current research interests include text and data mining, business intelligence and analytics, geographic information systems (GIS) and spatial analytics, and disaster planning and logistics.

Alan S. Abrahams is an Associate Professor in the Department of Business Information Technology, Pamplin College of Business, Virginia Tech. He received a Ph.D. 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 text mining for defect discovery. He is a Senior Editor of *Decision Support Systems* and has published in a variety of journals including *Production and Operations Management*, *Decision Support Systems*, *Expert Systems with Applications*, *Journal of Computer Information Systems*, *Communications of the AIS*, and *Group Decision and Negotiation*.

Please cite this article as: D.M. Goldberg, A.S. Abrahams, A Tabu search heuristic for smoke term curation in safety defect discovery, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.012