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