Automated extraction of product feedback from online reviews: Improving efficiency, value, and total yield

David Michael Goldberg

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Business Information Technology

Alan S. Abrahams, chair
Jason K. Deane
Cliff T. Ragsdale
Terry R. Rakes
Loren P. Rees

March 20, 2019
Blacksburg, VA

Keywords: text analytics, online reviews, business intelligence, heuristics, classification.

Automated extraction of product feedback from online reviews: Improving efficiency, value, and total yield

David Michael Goldberg

Academic abstract

In recent years, the expansion of online media has presented firms with rich and voluminous new datasets with profound business applications. Among these, online reviews provide nuanced details on consumers' interactions with products. Analysis of these reviews has enormous potential, but the enormity of the data and the nature of unstructured text make mining these insights challenging and time-consuming. This paper presents three studies examining this problem and suggesting techniques for automated extraction of vital insights.

The first study examines the problem of identifying mentions of safety hazards in online reviews. Discussions of hazards may have profound importance for firms and regulators as they seek to protect consumers. However, as most online reviews do not pertain to safety hazards, identifying this small portion of reviews is a challenging problem. Much of the literature in this domain focuses on selecting "smoke terms," or specific words and phrases closely associated with the mentions of safety hazards. We first examine and evaluate prior techniques to identify these reviews, which incorporate substantial human opinion in curating smoke terms and thus vary in their effectiveness. We propose a new automated method that utilizes a heuristic to curate smoke terms, and we find that this method is far more efficient than the human-driven techniques. Finally, we incorporate consumers' star ratings in our analysis, further improving prediction of safety hazard-related discussions.

The second study examines the identification of consumer-sourced innovation ideas and opportunities from online reviews. We build upon a widely-accepted attribute mapping

framework from the entrepreneurship literature for evaluating and comparing product attributes. We first adapt this framework for use in the analysis of online reviews. Then, we develop analytical techniques based on smoke terms for automated identification of innovation opportunities mentioned in online reviews. These techniques can be used to profile products as to attributes that affect or have the potential to affect their competitive standing. In collaboration with a large countertop appliances manufacturer, we assess and validate the usefulness of these suggestions, tying together the theoretical value of the attribute mapping framework and the practical value of identifying innovation-related discussions in online reviews.

The third study addresses safety hazard monitoring for use cases in which a higher yield of safety hazards detected is desirable. We note a trade-off between the efficiency of hazard techniques described in the first study and the depth of such techniques, as a high proportion of identified records refer to true hazards, but several important hazards may be undetected. We suggest several techniques for handling this trade-off, including alternate objective functions for heuristics and fuzzy term matching, which improve the total yield. We examine the efficacy of each of these techniques and contrast their merits with past techniques. Finally, we test the capability of these methods to generalize to online reviews across different product categories.

Automated extraction of product feedback from online reviews: Improving efficiency, value, and total yield

David Michael Goldberg

General audience abstract

This dissertation presents three studies that utilize text analytic methods to analyze and derive insights from online reviews. The first study aims to detect distinctive words and phrases particularly prevalent in online reviews that describe safety hazards. This study proposes algorithmic and heuristic methods for identifying words and phrases that are especially common in these reviews, allowing for an automated process to prioritize these reviews for practitioners more efficiently. The second study extends these methods for use in detecting mentions of product innovation opportunities in online reviews. We show that these techniques can used to profile products based on attributes that differentiate them from competition or have the potential to do so in the future. Additionally, we validate that product managers find this attribute profiling useful to their innovation processes. Finally, the third study examines automated safety hazard monitoring for situations in which the yield or total number of safety hazards detected is an important consideration in addition to efficiency. We propose a variety of new techniques for handling these situations and contrast them with the techniques used in prior studies. Lastly, we test these methods across diverse product categories.

Acknowledgements

In my seven years at Virginia Tech, I have been fortunate to be surrounded by world-class teachers, mentors, and collaborators that have guided and believed in me unfailingly. My five committee members – Alan Abrahams, Jason Deane, Cliff Ragsdale, Terry Rakes, and Loren Rees – have been central to my journey. They have been the finest teachers, most caring mentors, and most thoughtful collaborators I could ever have hoped for. Thank you to each for the years of generosity, kindness, and support.

I am also thankful to have had the opportunity over the years to attend classes taught by Mark Sheldon, Stéphane Collignon, Bill Carstensen, and Yang Shao, each of whom encouraged me along the way.

My family has always supported and believed in me through the years. Thank you to my two proud parents – affectionately Kebbie and Umsley – for their encouragement every step of the way. Thank you also to Patricia, who tolerated numerous long-winded explanations of my latest projects these past few years. She has been a patient, devoted, and loving companion and always encouraged me never to be anything less than my best.

Thank you to the friends that I made along the way – Nohel, Sukhwa, Rich, Zach, and David 2 – for all the help and support. It has been an honor working, commiserating, and best of all laughing together.

Last but certainly not least, I owe an enormous debt of gratitude to my feline friends. Thank you to Oreo, whose quest to keep my lap warm has endured for nine lives, and to Trouble, who ensured that I could never work too long at a time without stopping to admire the finer things in life.

Note

The work presented in the first chapter of this dissertation is published as follows.

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

Table of contents

Introduction

The spread of Internet connectivity in recent years has brought about massive changes in business intelligence. IHS [22] estimates that humanity currently operates over 20 billion Internet-connected devices across the world, and they estimate that this figure will increase to about 75 billion devices by 2025. Coinciding with the spread of Internet connectivity, online word-of-mouth (WOM), or the informal exchange of information from person to person, has also spread exponentially [28]. Today, Internet users are able to share user-generated content on social media sites (Facebook, Twitter, etc.), web forums/discussion boards, online review platforms, and even through multimedia formats (YouTube, Vimeo, etc.). The growth of this online content represents an incredible opportunity for business intelligence: firms that can harness the Internet to understand consumer preferences and capitalize upon them stand to improve their competitive positions substantially. Whereas in prior decades, firms relied on costly methods such as consumer surveys and focus groups to source business intelligence information [35], much of that feedback is now available online. In recent years, online reviews have been a particularly fruitful source of business intelligence [10, 11, 14, 21, 32], as firms can garner insights from feedback that is posted with the specific intent of evaluating their products.

Despite the promise of online media and online reviews in particular as a source for business intelligence, these data sources are not without limitations and difficulties. As much as the volume of online feedback serves to offer an enormous sample of consumer insights, it can be challenging for firms to extract the most relevant information from this deluge of data. Information systems literature dating back to the 1980s acknowledges the problem of "information overload" [12, 19]. Hemp [18] describes this phenomenon as a blurring of the line between worthwhile information and distracting information; as firms attempt to manage a

copious volume of data, they may struggle to separate the crucial signals of actionable information from the overwhelming noise. As such, prioritizing these datasets is a key step for firms attempting to make sense of online WOM. The information retrieval field focuses somewhat on these types of research problems [15, 34], although applications of information retrieval often slant more towards computer science than business analytics, such as search engines [9] or probabilistic relevance models [7, 36]. Online reviews are a popular source of this type of data, as they provide feedback specifically tailored to a firm's product lines. However, the unstructured and voluminous nature of online reviews raises difficulties in pursuing large-scale analyses of consumer feedback.

One aim of online reviews is to retrospectively assess how well each product has performed in practice. Firms may have a variety of concerns about their products' performance, such as safety hazards, which are especially significant concerns because they carry the risk of injury and/or death for consumers and may lead to subsequent lawsuits and/or recalls [17, 25]. Product use cases in practice often differ from product testing, making some safety hazards difficult to predict in advance of product release [38]. As a result, detecting defective products is both valuable and essential. Abrahams et al. [1] propose SMART, an integrated text analytics framework for product defect detection in online media. The authors argue that text analytic methods must be specifically tuned to domains of interest to produce efficacious results. This framework and methodology have been applied broadly to numerous industries [1-4, 23, 33, 41], although the characteristics of each industry represent complex linguistic challenges, and existing techniques are not equally effective in all applications.

More broadly, firms may wish to pursue an understanding of which attributes of their products may be perceived as positive or negative. This step allows firms to understand their

competitive position within their industry by comparing their products to competitors. As monitoring these reviews is an effort to understand consumer perceptions of products, the attribute mapping framework described by MacMillan and McGrath [26, 27] offers a useful starting point for prioritizing information. The authors distinguish between positive, negative, and neutral consumer sentiment about each product attribute; further, the authors subdivide these categories based upon consumers' likelihood to make purchasing decisions derived from their initial sentiment. Yet, identifying and mapping these attributes can be a challenge in itself, as consumers' perceptions of products often differ substantially from those products' designers [31]. Engaging with thorough and real-time consumer feedback offers firms an opportunity to gather this key form of intelligence and crystallize their understanding of their products' competitive positions.

This dissertation seeks to propose and adapt new text analytic and machine learning methods for extracting critical information from online reviews. It offers both methodological and theoretical contributions. Methodologically, the first and third chapters of the dissertation propose new methods for prioritizing online content to extract the most relevant information. Each study applies these methods to the detection of safety hazards in online reviews, as these represent particularly severe instances of product defects [17, 25, 38]; however, the techniques are also broadly applicable for other information extraction efforts. The first study presents a novel technique for prioritizing online reviews for firms that only have the resources or capacity to manually assess a small portion, while the third study instead approaches the problem by also considering the total number of true positives detected. The third study unifies the different types of algorithmic approaches proposed and offers guidance as to the circumstances in which to deploy each technique. The second study aims to offer an empirical validation of the

aforementioned attribute mapping framework through analyses of online reviews by focusing on extracting information pertaining to product innovation opportunities; it also extends that framework to reconcile the characteristics of the online review format. This framework allows firms to understand their competitive position and to source consumer-driven feedback from online reviews that offers actionable improvements to their product offerings.

Study 1: Improving *efficiency* of safety hazard monitoring in online reviews

The first study extends prior work to offer a revised method for identifying the presence of safety hazards in online reviews. Safety hazards represent a particularly pressing form of product defect, and instances can result in extremely expensive product recalls [17, 25, 38]. Although techniques to identify these online reviews of interest are highly studied [1-4, 41], the existing text mining methodology has relied greatly on human judgment to identify terms that delineate safety hazards ("smoke terms"). These methods aim to aid with **prioritization**, or a ranking of online reviews from most likely to least likely to contain the target classification. This study proposes a new method in which this subjective manual process is replaced with an automated Tabu search algorithm that maximizes the number of true positives found in the top-ranking reviews. The two methods are compared, and the revised method offers statistically significant improvement over a large sample of human-curated smoke term lists. Surprisingly, the study finds that shorter and more targeted smoke term lists outperform the longer smoke term lists chosen by manual curators. Additionally, unlike prior work, this study incorporates online review star ratings into the detection of safety hazards, which provides further statistically significant improvement. In total, the new technique approximately doubles the performance of prior techniques.

Research on detecting features of interest in online reviews by using supervised machine learning models is quite common, particularly in detection of safety hazards [1-4, 41]. As these models form the basis of many new research efforts, using more effective analytic techniques will offer more performance and credence to future research methods that may employ this model. This model's applications are not limited to purely safety hazard detection. For example, in the second study, this model will be extended to focus on product development applications,

identifying feature requests and other innovation-related discussion in online media. Researchers may be able to extend this model to many text analytics applications. The proposed methodology is also rather unique in that it offers an application of a management science heuristic (Tabu search) to the text analytics domain. These two domains are often quite distinct, and it is unusual to apply a method from one domain in the context of the other. Yet, the study demonstrates the applicability of heuristics in a text mining context to improve machine learning performance, and hopefully it will inspire future work to explore applications of management science in text mining.

This study has many benefits for organizations. Many organizations have noted the importance of detecting safety hazards in their products as early as possible. Safety-related product recalls are often enormously expensive, with recent recalls reaching billions of dollars in costs to firms in replacing defective products, settling class action suits, and paying federal penalties [17, 25]. Sifting through customer feedback to detect these hazards as quickly as possible is paramount, but it is also difficult given the immense volume of online feedback [8]. Methods that allow for the prioritization of this content alleviate a major labor requirement in constant monitoring of online media and provide expedient feedback that may quickly identify hazards and save lives. These monitoring efforts should culminate in **remediation** or acting upon online feedback to remedy products. Most directly, the study establishes sets of smoke terms that may be used to identify safety hazards in the countertop appliances and over-the-counter medicine industries. For practitioners in these industries, these smoke term lists represent the best available methods for quickly monitoring online media and detecting potential safety hazards. More broadly, this study establishes a method for training high-performing smoke term lists in any industry, and the method offers superior precision compared to prior work.

Study 2: Delivering business *value* through rapid identification of innovation opportunities in online reviews

The second study proposes to extend prior work on automated defect detection in online media to product innovation opportunities. In prior work (e.g., [1, 3]), automated detection efforts have been focused largely on safety and performance defects. This study engages with the attribute mapping framework proposed by MacMillan and McGrath [29], which delineates between various product development opportunities, such as "differentiators" that distinguish superior products and "enragers" over which consumers become irritated with poor quality products. Using the heuristic methods for text prioritization proposed in the first study for effectively classifying online reviews, this study will seek to automatically identify these innovation opportunities in online reviews. The results of this study are verified by a blind assessment of the usefulness of the online media extracted by senior-level managers from a large Fortune 1000-listed manufacturer of countertop appliances earning over $500 million in revenue per year.

The second study is novel with respect to two different areas of research. First, in the text analytics arena, studies that examine users' innovation-related feedback are quite limited. Lee and Bradlow [24] develop a text analytic model focused on extracting marketing data from online reviews. The authors' technique identifies broad marketing trends at the aggregate level, such as which known product features are mentioned most often in online reviews. The authors argue that it is necessary for researchers to further explore **user needs**, or product attributes that deliver value to consumers, as it is still difficult for firms to expeditiously determine consumers' precise preferences and requirements. Second, this study proposes an empirical validation of MacMillan and McGrath's attribute mapping framework, which offers a theoretical view of

product innovation opportunities [26, 27], but the evidence of the framework's effectiveness has been anecdotal or in the form of case studies rather than a rigorous empirical validation [5, 30, 39]. The proposed study would both be a novel use of state-of-the-art text mining methodology in a seldom-explored arena and a vital empirical validation of a long-standing theory in the literature.

Product designers often consider many data sources when revising their products, including focus groups, customer complaints and warranty claims, and their own ingenuity [35]. While not all customer suggestions are feasible, experimental evidence suggests that product design that considers customer feedback out-sells product design performed in laboratory [31]. In fact, various academic outlets have called for methods that emphasize consumer-driven innovation [6, 13]. In this sense, the immense volume of timely product feedback in the form of online product reviews represents a great and untapped opportunity. While many practitioners surely make use of online reviews in their product development processes, online reviews are so voluminous that it is nearly impossible for practitioners to systematically read them all [8]. This study emphasizes prioritization: which reviews are the most pertinent to the feedback category of interest (safety hazards, irritators, feature requests, or compliments). Using these techniques, practitioners can narrow the process of soliciting feedback from online reviews down to a smaller and more manageable sub-sample of useful information.

Study 3: Maximizing *total yield* in safety hazard monitoring of online reviews

The third study proposes differently motivated techniques for the curation of smoke terms in online review analysis. The Tabu search text analytic technique proposed in the first study effectively ensures that online reviews are prioritized, so top-ranking reviews are extremely likely to be true positives. However, a limitation of this study is in the **depth** of its solutions, or the extent to which each solution maximizes the total number of true positives detected. Each smoke term list offers powerful predictions for a few hundred reviews, but as the smoke term lists tend to be rather short, they are most useful in prioritizing that top portion. In other words, the technique is designed to maximize precision, but it does not consider recall. The difficulty of balancing precision and recall is often acknowledged in machine learning classification research [16, 40]. In safety hazard detection, identifying as many of the target classification as possible is a serious concern, particularly for industries in which hazards are especially severe [20]. In the third study, consideration is given to the case of firms that may also wish to consider depth as well as precision, ensuring that as many true positive reviews as possible are accurately identified. For firms with the capacity for greater levels of manual examination, this would provide a more useful tool. Additionally, for firms in industries for which false negatives are especially costly, this adaptation would be a necessity. This study proposes several new prospective methods for addressing this problem. First, the study proposes "piecewise" smoke terms, or several consecutive iterations of the Tabu search heuristic that maximize precision over different parts of the distribution. Second, it proposes minimizing the ranks of true positives in the curation dataset, effectively shifting all true positive reviews toward the top of the distribution. Third, it proposes fuzzy string matching, which increases the number of reviews spanned by each smoke term. Finally, the study explores the possibility of generating cross-

category smoke terms, which can be immediately applied to any product category if a more category-specific smoke term list has yet to be developed.

Similar to the first study, the third study explores the popular area of extracting meaning from online reviews, with a particular focus on the detection of safety hazards [1-4, 41]. In this research community, the third study will present several new potential methods for prioritizing online reviews. Although safety hazards are used as the study area of interest in the study, applications are wide-ranging, and the same technique can be used for other areas in which researchers wish to extract as many true positives as possible from textual data. For example, some research has examined using text analytic methods to classify terror chatter [37], a domain in which it is vital not only to prioritize a huge volume of data but also to ensure that as many true positives as possible are detected. For organizations, especially in industries for which safety hazards are severe [20], these techniques provide a rapid means of sourcing vital feedback from real-time crowdsourced intelligence. Identifying product safety hazards as soon as possible allows firms to start remediating, which mitigates the expensive and lengthy product recall process for defective products [17, 20, 25, 38].

Research framework

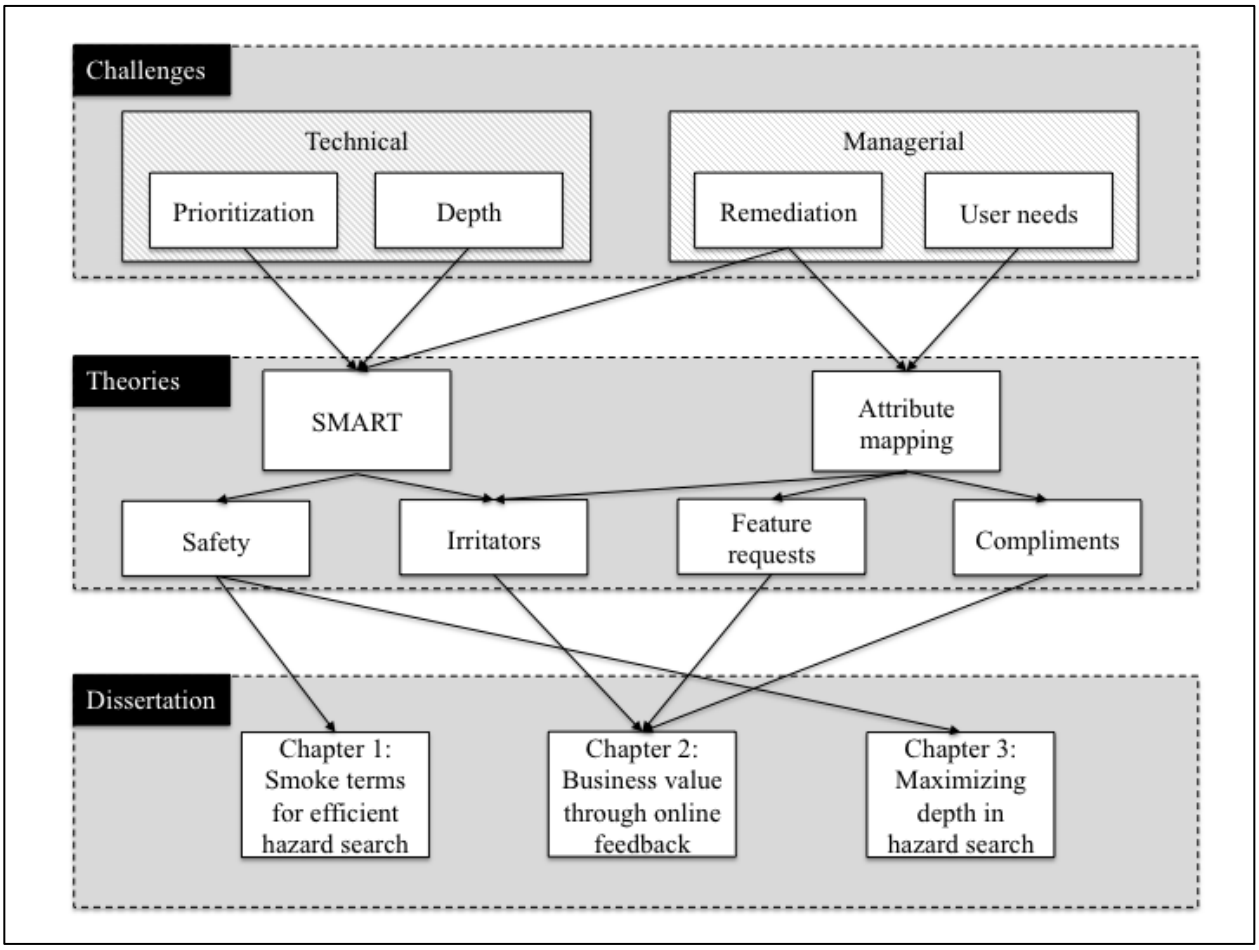

Figure 1. Proposed research framework.

Figure 1 provides a depiction of the research framework proposed in this dissertation. The proposed studies address two broad types of challenges: technical challenges (algorithms, techniques, and implementations) and managerial challenges (making informed business decisions). Each work in the dissertation addresses both types of challenges for pressing issues in academia and industry. First, what is an effective technical solution to a difficult problem? Second, how can researchers or practitioners use these insights to improve their work? The dissertation contextualizes these challenges in prior work by engaging with SMART [1], which suggests a framework for studying the domain-specific challenges in online reviews, and the

attribute mapping framework [26, 27], which provides managers with a useful tool for contextualizing consumer feedback and prioritizing their handling of product attributes. SMART has specifically suggested means for addressing safety hazards and performance defects. The first and third studies study safety hazard detection explicitly and explore methodological solutions for detecting these concerns more rapidly or completely. The second study also discusses identification of irritators, or specific performance-related complaints. Furthermore, the second study examines each of the technical challenges in a managerial context, and it illustrates how managers may take advantage of these solutions to better understand and improve upon their product offerings. Taken together, these three works should provide researchers and practitioners with a powerful new toolkit for interpreting and capitalizing upon online reviews.

References

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

[2] 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, Decision Support Systems, 55(4) (2013) 871-882.

[3] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decision Support Systems, 54(1) (2012) 87-97.

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

[5] R. Amit, C. Zott, Value creation in e-business, Strategic Management Journal, 22(6-7) (2001) 493-520.

[6] L.A. Bettencourt, A.W. Ulwick, The customer-centered innovation map, Harvard Business Review, 86(5) (2008) 109.

[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research, 3(Jan) (2003) 993-1022.

[8] BrightLocal, Local Consumer Review Survey, in, (2016).

[9] S. Büttcher, C.L. Clarke, G.V. Cormack, Information retrieval: Implementing and evaluating search engines, (MIT Press, 2016).

[10] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: Online book reviews, Journal of Marketing Research, 43(3) (2006) 345-354.

[11] C. Dellarocas, N. Awad, X. Zhang, Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning, ICIS 2004 Proceedings, 30 (2004) 379-386.

[12] P.J. Denning, ACM president's letter: Electronic junk, Communications of the ACM, 25(3) (1982) 163-165.

[13] K.C. Desouza, Y. Awazu, S. Jha, C. Dombrowski, S. Papagari, P. Baloh, J.Y. Kim, Customer-driven innovation, Research-Technology Management, 51(3) (2008) 35-44.

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

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

[16] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3(Mar) (2003) 1289-1305.

[17] Y. Hagiwara, T. Taniguchi, Takata puts worst-case airbag recall costs at $24 billion, in: Bloomberg, (2016).

[18] P. Hemp, Death by information overload, Harvard Business Review, 87(9) (2009) 83-89.

[19] S.R. Hiltz, M. Turoff, Structuring computer-mediated communication systems to avoid information overload, Communications of the ACM, 28(7) (1985) 680-689.

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

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

[22] IHS, The Internet of Things: A movement, not a market, in, (London, 2017).

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

[24] T.Y. Lee, E.T. Bradlow, Automated marketing research using online customer reviews, Journal of Marketing Research, 48(5) (2011) 881-894.

[25] Y. Lee, Samsung Note 7 recall to cost at least $5.3 billion, in: Associated Press, (2016).

[26] I.C. MacMillan, R.G. McGrath, Discover your products' hidden potential, Harvard Business Review, 74(3) (1996) 58-73.

[27] I.C. MacMillan, R.G. McGrath, Discovering new points of differentiation, Harvard business Review, 75 (1997) 133-145.

[28] J. McAuley, R. Pandey, J. Leskovec, Inferring networks of substitutable and complementary products, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM, 2015), pp. 785-794.

[29] R.G. McGrath, I.C. MacMillan, The entrepreneurial mindset: Strategies for continuously creating opportunity in an age of uncertainty, (Harvard Business Press, 2000).

[30] R.G. McGrath, A. Nerkar, Real options reasoning and a new look at the R&D investment strategies of pharmaceutical firms, Strategic Management Journal, 25(1) (2004) 1-21.

[31] D.L. Meadows, Estimate accuracy and project selection models in industrial research, Industrial Management Review, 9(3) (1968) 105.

[32] S.M. Mudambi, D. Schuff, What makes a helpful review? A study of customer reviews on Amazon.com, MIS Quarterly, 34(1) (2010) 185-200.

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

[34] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, 2(1-2) (2008) 1-135.

[35] J. Pruitt, T. Adlin, The persona lifecycle: keeping people in mind throughout product design, (Elsevier, 2010).

[36] N. Rekabsaz, M. Lupu, A. Hanbury, G. Zuccon, Generalizing translation models in the probabilistic relevance framework, in: Proceedings of the 25th ACM international on conference on information and knowledge management, (ACM, 2016), pp. 711-720.

[37] E. Riloff, W. Lehnert, Information extraction as a basis for high-precision text classification, ACM Transactions on Information Systems, 12(3) (1994) 296-333.

[38] N.G. Rupp, The attributes of a costly recall: Evidence from the automotive industry, Review of Industrial Organization, 25(1) (2004) 21-44.

[39] D.G. Sirmon, M.A. Hitt, Managing resources: Linking unique resources, management, and wealth creation in family firms, Entrepreneurship Theory And Practice, 27(4) (2003) 339-358.

[40] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research, 2(Nov) (2001) 45-66.

[41] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decision Support Systems, 90 (2016) 23-32.

Chapter 1: Improving efficiency of safety hazard monitoring in online reviews

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.

Keywords: text mining, online reviews, Tabu search, heuristics, defects, business intelligence

1.0 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 [25], while the recall of Takata airbags in the automotive industry was estimated to cost the firm up to $24 billion [18]. 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 [33].

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 [33], and the prevalence of product recalls at over $1 trillion of total costs in the United States each year [11] 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 [3]. Firms may be pleased that a minority of online reviews refer 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 [8, 9, 20], 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. [1-3] 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 [1-3]. Beyond this initial work in the automotive industry, further research by Winkler et al. [38], Law et al. [24], and Adams et al. [4] 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 techniques such as those proposed in Fan et al. [15] 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 [1-4, 24, 38]. 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 [12]. 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 [31]. Recalls in the OTC medicine industry are also problematic, such as a 2016 nationwide recall of potentially lethal children's medications [28]. 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 [1, 4, 38], 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.

2.0 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 [9]. WOM communication in online reviews includes vital information on those consumers' perceptions of product quality [20], and these reviews have further impacts upon future consumption of those products by other consumers reading the reviews [9]. 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 [20].

Chevalier and Mayzlin [9] found evidence that the mean star rating in product reviews was positively related with the subsequent sales of associated products, while Duan et al. [14] 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. [27] 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 [14, 40]. Researchers extract text from online sources, such as product reviews [1-4, 9, 20, 24, 38] and social media [6] 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 [30], ANEW [7], and the Harvard General Inquirer [22]. Some sentiment analysis techniques such as SentiStrength [37] 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 [35]. Tang et al. [36] provide a comprehensive overview of prior sentiment analysis literature in online reviews. Sentiment analysis has even been used to predict stock performance at the firm [40] and market levels [6].

Due to its ability to distinguish negative from positive opinions, sentiment analysis has been employed to detect product safety concerns [21, 39]. 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 [4, 24]. 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 [1-4, 24, 38]. 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 [1-4, 24, 38].

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 [1, 3]. 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 [32]. Fan et al. [15]'s pivotal work proposes several further innovative strategies for evaluating term relevance. First, Fan et al. [15] propose the Relevance Correlation Value (RCV), based on the Vector Space model [34], for defining a term's relevance based upon the number of times it appears in training documents classified as relevant. Fan et al. [15] also propose an adaptation upon Ng et al. [29]'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. [1, 2], although more recent works on defect discovery have utilized the CC score extensively and observed excellent performance [4, 24, 38]. 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 [1-4, 24, 38]. Indeed, research has observed a decline in the quality of terms after those ranked in the top few hundred [2, 38]. Researchers often set arbitrary cut-offs for the number of terms to include in

final smoke term lists or minimum relevance scores for inclusion [24, 38]; in addition, smoke term lists are manually filtered to retain only the terms intuitively believed to provide the greatest precision [1-4, 24, 38]. The literature has provided several rationales for the removal of terms from this initial list but acknowledges that this process is substantially subjective [1-4, 24, 38].

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 [1-4, 24, 38]. 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 [1-4, 24, 38]. 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 [38]. 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 [1-4, 24, 38]. 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 [38]. 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 [2, 3]. Finally, researchers may remove spuriously prevalent words subjectively identified as irrelevant [1-4, 24, 38].

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.

3.0 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 [1-4, 24, 38], the issue of the variability of the efficacy of these lists has not yet been analyzed in the literature. We provide details 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 [1, 4, 38], 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.

4.0 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 them for potential defects [3]. 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 [38] or 200-ranked reviews [4, 24] 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 [26]. 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" [3]. 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 from appliances, 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, 4,142 countertop appliance reviews were tagged multiple times, and amongst these instances, taggers were unanimous 3,986 times (96.2% of cases) and disagreed just 156 times (3.6% of cases). Additionally, 2,157 OTC medicine reviews were tagged multiple times, for which taggers were unanimous 2,110 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 [38]. 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 [10] value of 0.867. For OTC medicine, we observed 91.3% agreement between the authority tagger and other taggers and a Cohen's κ [10] value of 0.827. Landis and Koch [23] rate agreement in this range as "almost perfect", while Fleiss et al. [16] rate agreement in this range as "excellent".

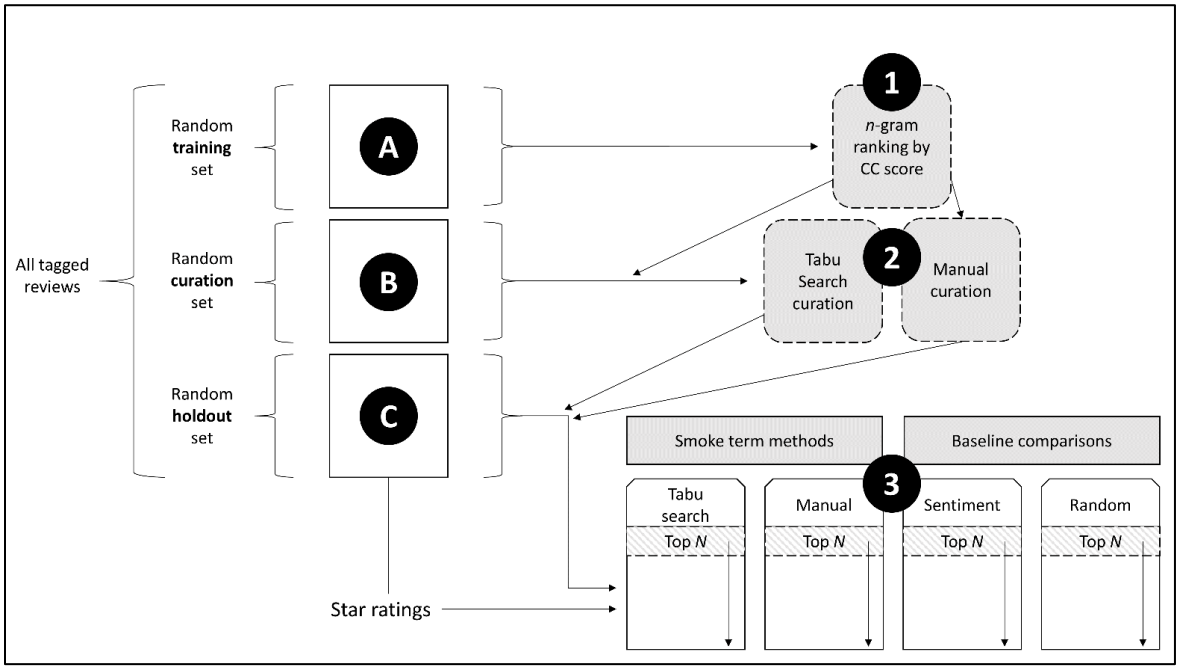

Figure 1. Proposed data processing steps.

In Figure 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. 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 [15]. 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 [1, 3, 4], 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 [4, 24, 38] has found the best performance when using the CC score proposed by Fan et al. [15]. 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 [1, 4, 24]. 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 3,624 OTC medicine reviews contained 9,709 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.

Table 1. Top-ranking *n*-grams from the training sample by CC score [15].

Panel A: Countertop appliances industry						
Rank	**Unigram**	**CC score**	**Bigram**	**CC score**	**Trigram**	**CC score**
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						
Rank	**Unigram**	**CC score**	**Bigram**	**CC score**	**Trigram**	**CC score**
1	studies	4,966.67	will i	6,083.75	my stomach is	5,212.88
2	rls	4,966.67	stomach pain	5,185.01	not something i	5,212.88
3	stone	4,966.67	milk of	4,966.67	to wear off	4,966.67
4	toll	4,966.67	leg syndrome	4,966.67	pain and cramping	4,966.67
5	capful	4,966.67	nasty i	4,966.67	afternoon i had	4,966.67
6	lack	4,966.67	the women	4,966.67	by afternoon i	4,966.67
7	magnesia	4,966.67	was fairly	4,966.67	never had this	4,966.67
8	enhance	4,966.67	of magnesia	4,966.67	just taking one	4,966.67
9	poisoning	4,966.67	restless leg	4,966.67	treatment for a	4,966.67
10	clog	4,966.67	strong so	4,966.67	maybe it was	4,966.67

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 [15]. 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 [15] 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 [1-4, 24, 38]. 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 minutes (5.95 minute standard deviation) curating each countertop appliances list and 13.30 minutes (5.55 minute 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 [15], we define a 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 $t = 1 \ldots T$, or $[x_1\ x_2\ x_3 \ldots\ x_T]$. 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 $i = 1 \dots m$, 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:

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

$$x_t \; binary \; \forall \; t$$

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 towards 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 [17]. 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 on 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 [1-4, 24, 38], 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 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*.

```
declare integer length ← 200 // Number of smoke terms to consider
declare array cutoffs ← [50, 100, 200] // Cutoffs considered (n(i))
declare array weights ← [0.333, 0.333, 0.333] // Weights considered (w(i))
declare array solution ← [ ]
declare array tabu_list ← [ ]

// Define initially empty smoke term list
for i = 1 to length
    solution.add(0)
end for
declare array best_solution ← solution

while not (stopping_condition()) // Time elapsed without improvement in objective

    declare array candidates ← [ ]
    declare array fitnesses ← [ ]

    for i = 1 to length

        // Generate adjacent solution
        declare array candidate ← solution
        if candidate[i] = 1 then
            candidate[i] ← 0
        Else
            candidate[i] ← 1
        end if

        // Evaluate fitness of smoke term list
        if not (candidate in tabu_list) then
            fitnesses.add(fitness(candidate, cutoffs, weights))
            candidates.add(candidate)
        end if

    end for

    declare integer index = argmax(fitnesses) // Determine best candidate
    solution ← candidates[index]
    tabu_list.add(solution)

    if fitnesses[index] > fitness(best_solution, cutoffs, weights) then
        best_solution ← solution // Update best solution
    end if

end while

return best_solution
```

The principle employed in lines 40-42 is an example of a greedy heuristic [17], 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 [13]. 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 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 [19] 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 [35, 36] and word frequency [1-4, 24, 38] 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. [27] 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 [27], a phenomenon also acknowledged by the psychology literature [5]. As star ratings are often predictive of product quality [20], 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 quality of products, while a

rating of 5 indicates the best 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.

5.0 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 [4, 24, 38]. 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.

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

Panel A: Countertop appliances industry (*N* = 102)						
Rank	**Unigram**	**Count**	**Bigram**	**Count**	**Trigram**	**Count**
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)						
Rank	**Unigram**	**Count**	**Bigram**	**Count**	**Trigram**	**Count**
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%)

"*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. 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.

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

Panel A: Performance of manually curated unigrams (*N* = 102).							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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).							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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).							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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

* indicates that scores were augmented by star ratings.

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

Panel A: Performance of manually curated unigrams (*N* = 139).							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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).							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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).							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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

* indicates that scores were augmented by star ratings.

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.

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%)

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

Panel A: Countertop appliances industry			
Unigram	**CC score**	**CC score rank**	**Manual curation rank**
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			
Unigram	**CC score**	**CC score rank**	**Manual curation rank**
capful	4,966.67	5	45
caution	4,626.46	13	5
liver	4,507.69	17	9
potentially	3,990.57	27	10
surrounding	3,990.57	28	160
caused	3,567.76	32	11
catastrophic	3,511.48	105	14
overdosed	3,511.48	109	6
chills	3,511.48	120	47
monster	3,511.48	195	104
heaving	3,511.48	200	38

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

Panel A: Countertop appliances industry			
Bigram	**CC score**	**CC score rank**	**Manual curation rank**
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
youre 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			
Bigram	**CC score**	**CC score rank**	**Manual curation rank**
stomach pain	5,185.01	2	1
caution if	4,966.67	15	5
by afternoon	4,966.67	18	124
it caused	4,966.67	19	12
with caution	4,612.61	48	6
periods of	3,990.57	79	67
that evening	3,990.57	90	96
severely burnt	3,511.48	147	24
your belly	3,511.48	166	146

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 [2, 38], 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.

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

Panel A: Countertop appliances industry			
Trigram	**CC score**	**CC score rank**	**Manual curation rank**
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			
Trigram	**CC score**	**CC score rank**	**Manual curation rank**
my stomach is	5,212.88	1	17
pain and cramping	4,966.67	4	1
wouldnt use it	4,966.67	16	16
going to faint	4,966.67	19	17
had this problem	4,966.67	31	21
and very bad	4,966.67	37	23
with caution if	4,966.67	38	10
it with caution	4,966.67	42	11
this product contains	3,990.57	57	25
sweats and flu	3,511.48	90	15
all day now	3,511.48	158	143
morning and still	3,511.48	171	135

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-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".

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

Panel A: Countertop appliances industry							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
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							
		Number of safety defects found in the top *N*-ranked reviews					
	List length	Top 50	Top 100	Top 200	Top 50*	Top 100*	Top 200*
Unigrams	11	15	21	32	20	28	43
Bigrams	9	14	21	30	15	24	38
Trigrams	12	14	20	29	17	25	42

* indicates that scores were augmented by star ratings.

"*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 [1-4, 24, 38]. 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.

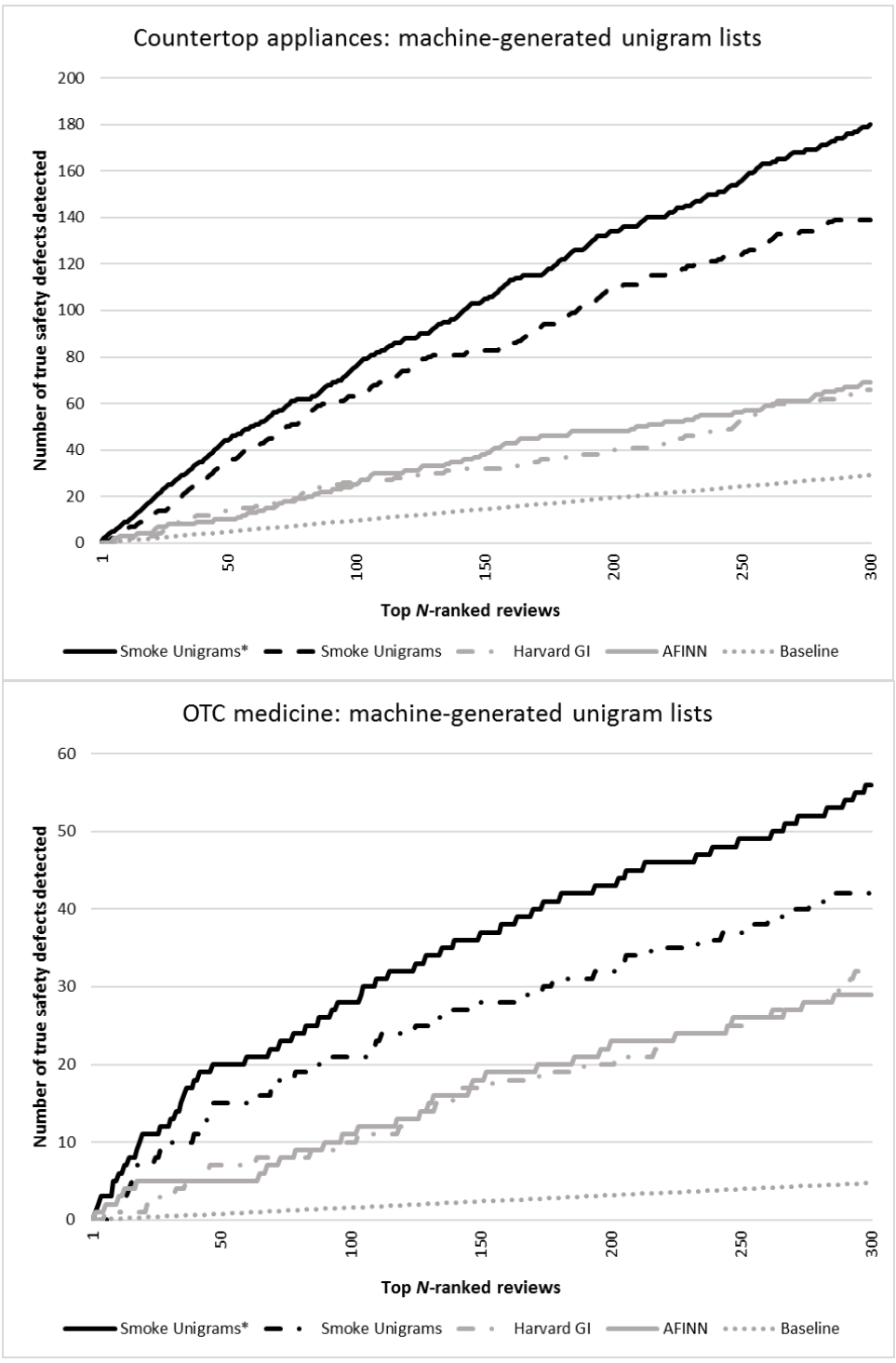

Figure 3. Performance of machine-curated unigram lists.

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 Figures 3-5, 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, 30] 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.

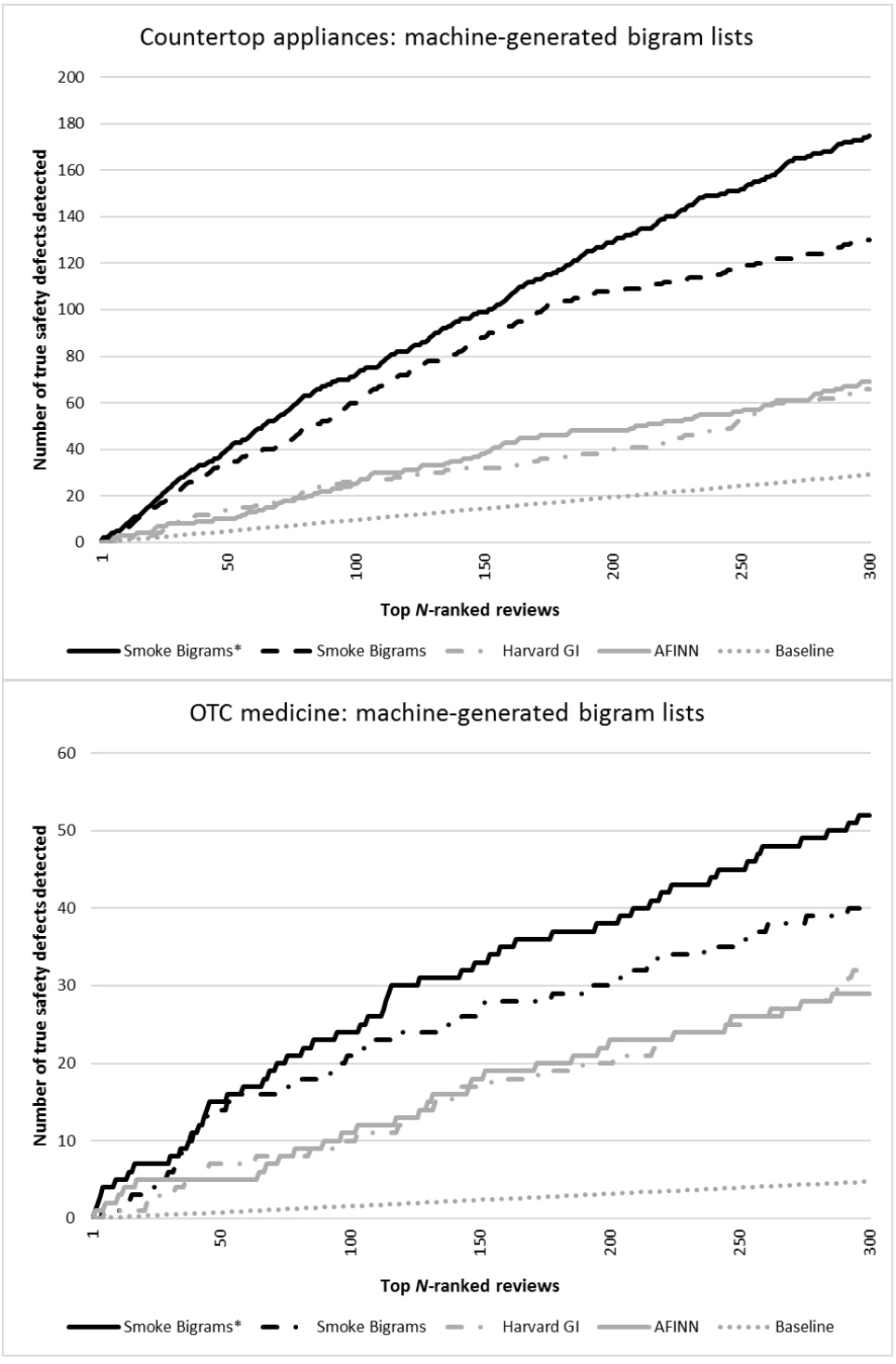

Figure 4. Performance of machine-curated bigram lists.

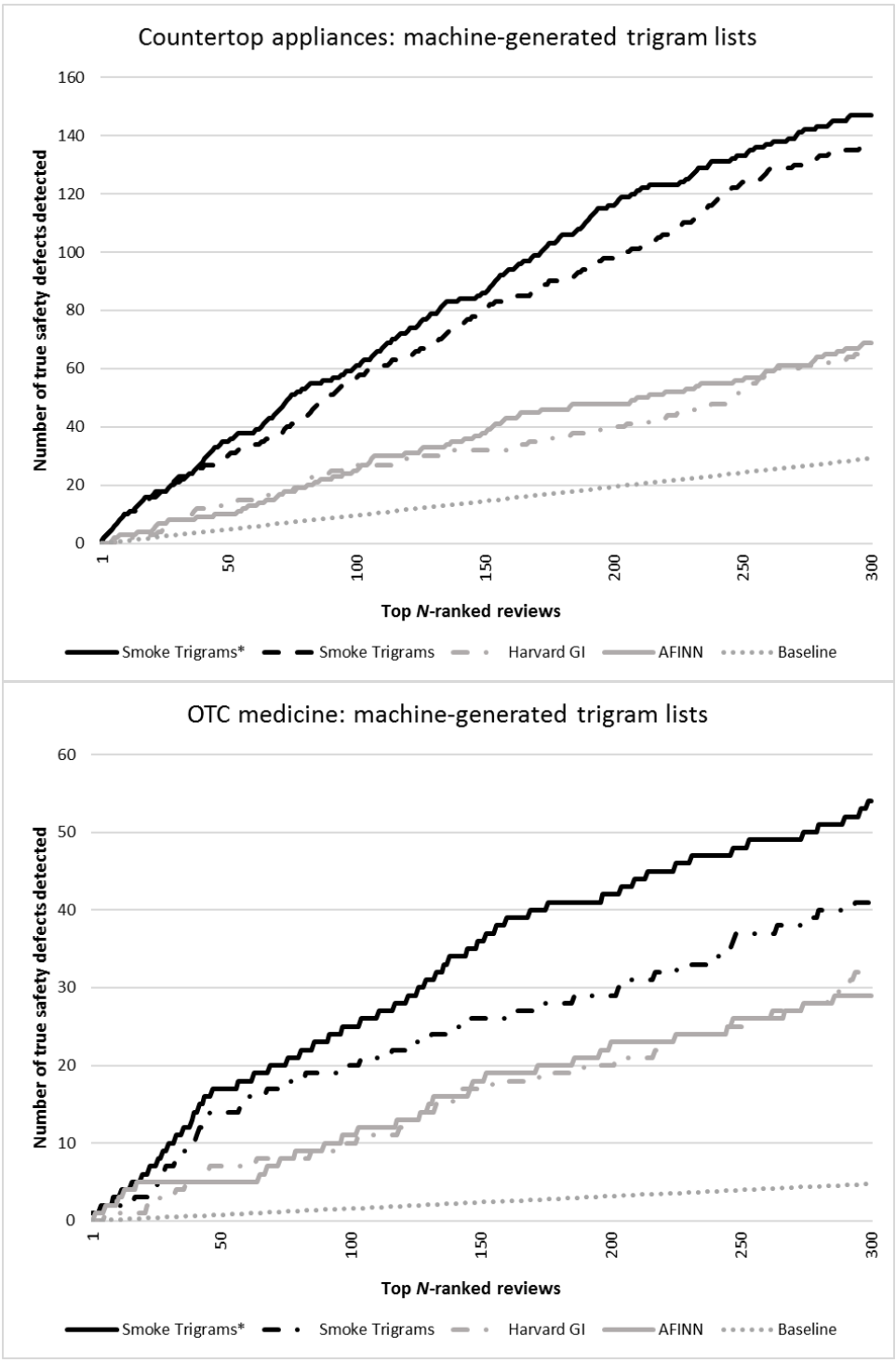

Figure 5. Performance of machine-curated trigram lists.

6.0 Limitations

The process of tagging reviews used in this paper required an ample level of human effort, and tagging methods have become standard in defect discovery literature for the development of training data [1-4, 24, 38]. 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]. We have no reason to expect that online reviews represent a biased estimate of safety defect prevalence in consumer products, but 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.

7.0 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 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 [1-4, 24, 38]; 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 [1-4, 24, 38], 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.

References

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

[2] 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, Decision Support Systems, 55(4) (2013) 871-882.

[3] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decision Support Systems, 54(1) (2012) 87-97.

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

[5] T.M. Amabile, Brilliant but cruel: Perceptions of negative evaluators, Journal of Experimental Social Psychology, 19(2) (1983) 146-156.

[6] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science, 2(1) (2011) 1-8.

[7] M.M. Bradley, P.J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, in, (Citeseer, 1999).

[8] BrightLocal, Local Consumer Review Survey, in, (2016).

[9] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: Online book reviews, Journal of Marketing Research, 43(3) (2006) 345-354.

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

[11] C.P.S. Commission, 2015 Annual Report, in, (2015).

[12] ConsumerReports, Appliance fires pose a safety concern, in, (2012).

[13] R. Dechter, J. Pearl, Generalized best-first search strategies and the optimality of A*, Journal of the ACM, 32(3) (1985) 505-536.

[14] W. Duan, B. Gu, A.B. Whinston, The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry, Journal of Retailing, 84(2) (2008) 233-242.

[15] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decision Support Systems, 40(2) (2005) 213-233.

[16] J.L. Fleiss, B. Levin, M.C. Paik, Statistical methods for rates and proportions, (John Wiley & Sons, 2013).

[17] F. Glover, Future paths for integer programming and links to artificial intelligence, Computers & Operations Research, 13(5) (1986) 533-549.

[18] Y. Hagiwara, T. Taniguchi, Takata Puts Worst-Case Airbag Recall Costs at $24 Billion, in: Bloomberg, (2016).

[19] J.H. Holland, Adaptation in natural and artificial systems, (University of Michigan Press, 1975).

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

[21] H. Isah, P. Trundle, D. Neagu, Social media analysis for product safety using text mining and sentiment analysis, in: 14th UK Workshop on Computational Intelligence, (IEEE, 2014), pp. 1-7.

[22] E.F. Kelly, P.J. Stone, Computer recognition of English word senses, (North-Holland, 1975).

[23] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics, 33(1) (1977) 159-174.

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

[25] Y. Lee, Samsung Note 7 recall to cost at least $5.3 billion, in: Associated Press, (2016).

[26] J. McAuley, R. Pandey, J. Leskovec, Inferring networks of substitutable and complementary products, in: Proceedings of the 21th SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM, 2015), pp. 785-794.

[27] S.M. Mudambi, D. Schuff, Z. Zhang, Why aren't the stars aligned? An analysis of online review content and star ratings, in: 47th Hawaii International Conference on System Sciences, (IEEE, 2014), pp. 3139-3147.

[28] D.J. Neal, 3 children’s medicines recalled because of potentially lethal ingredient, in: Sacramento Bee, (2016).

[29] H.T. Ng, W.B. Goh, K.L. Low, Feature selection, perceptron learning, and a usability case study for text categorization, in: ACM SIGIR Forum, (ACM, 1997), pp. 67-73.

[30] 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) 93-98.

[31] B. Perlow, 22 models of Cuisinart food processors recalled after reports of blade breaking off, in: ABC News, (2016).

[32] S.E. Robertson, On relevance weight estimation and query expansion, Journal of Documentation, 42(3) (1986) 182-188.

[33] N.G. Rupp, The attributes of a costly recall: Evidence from the automotive industry, Review of Industrial Organization, 25(1) (2004) 21-44.

[34] G. Salton, The SMART retrieval system—experiments in automatic document processing, (Prentice Hall, 1971).

[35] S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior, Journal of Management Information Systems, 29(4) (2013) 217-248.

[36] H. Tang, S. Tan, X. Cheng, A survey on sentiment detection of reviews, Expert Systems with Applications, 36(7) (2009) 10760-10773.

[37] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, Journal of the American Society for Information Science and Technology, 63(1) (2012) 163-173.

[38] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decision Support Systems, 90(2016) 23-32.

[39] C.C. Yang, H. Yang, L. Jiang, M. Zhang, Social media mining for drug safety signal detection, in: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, (ACM, 2012), pp. 33-40.

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

[41] G.K. Zipf, The psycho-biology of language, (Houghton Mifflin, 1935).

Appendix A: Supplementary material

In this supplement, we describe the scoring metric employed by the smoke term methodology. The smoke term methodology is discussed variously in [5, 9-13]; each work describes the process for smoke term scoring, although none of the works defines the methodology in terms of formal mathematical equations or matrix operations. As such, in this supplement, we establish a formal mathematical description for this scoring metric. In addition to our discussion, we demonstrate the scoring metric on an example dataset.

The first step in the methodology involves using an information retrieval technique such as the CC score [14] to establish initial candidate smoke terms. The CC scoring algorithm assigns each term a value based on its document relevance, or the frequency of the term in relevant documents compared to the frequency of the term in irrelevant documents. Higher scores indicate terms that are better predictors of relevant documents, as these terms occur very frequently in relevant documents and very infrequently in irrelevant documents.

Let i denote the identifier of each term in the smoke term dictionary, and let j denote the identifier of each review. We define D as a row vector representing the candidate smoke term dictionary. We denote each term in the dictionary as d_i, where the terms are indexed from 1 ... k. k represents the total number of unique terms in the dictionary. Thus, $D = [d_1 \quad d_2 \quad \dots \quad d_k]$. Correspondingly, we also define C as a row vector, also indexed from 1 ... k. Each entry in C represents the CC score for the i^{th} term in the smoke term dictionary, or the weight applied to the i^{th} term. We denote each CC score as c_i, again indexing the vector from 1 ... k. Thus, $C = [c_1 \quad c_2 \quad \cdots \quad c_k]$. Consider the following example vectors for D and C.

$$D = [\text{"}Dangerous\text{"} \quad \text{"}Fire\text{"} \quad \text{"}Hazard\text{"}]$$

$$C = [150 \qquad 100 \qquad 50]$$

Next, we define M as the document-term matrix relating the terms in the smoke term dictionary to the reviews in the dataset. A document-term matrix captures the frequency at which each term occurs in each document, or in this case, the frequency at which each smoke term occurs in each review. The columns of M are indexed by each of the terms in D, or from $1 \ldots k$. The rows of M are indexed by each of the reviews in the dataset, or from $1 \ldots r$. M is thus an $r \times k$ matrix, with r rows (one row for each review) and k columns (with the i^{th} column in M corresponding to the i^{th} term in the smoke term dictionary, D). As safety defect reports are relatively rare, we find that most smoke terms do not occur in most reviews. Therefore, M tends to be a sparse matrix in which most values are zero. Consider the following example matrix for M.

		$\boldsymbol{i = 1}$	$\boldsymbol{i = 2}$	$\boldsymbol{i = 3}$
j	**Review text**	**"Dangerous" frequency**	**"Fire" frequency**	**"Hazard" frequency**
1	This product was excellent! Highly recommend.	0	0	0
2	A sharp piece broke off the side, which poses a hazard.	0	0	1
3	Didn't work at all! Can't believe they still sell these.	0	0	0
4	Worked OK, but nothing special.	0	0	0
5	Caught on fire instantly. Fire almost burned my house down.	0	2	0

$$M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}$$

Finally, we define Y as the column vector of smoke term scores for each review. We define each review's smoke term score as y_j indexed from $1 \ldots r$. Thus, $Y = [y_1 \quad y_2 \quad \cdots \quad y_r]$. Y is calculated as $MC^\top$, which increments each review's smoke term score by c_i each time that d_i occurs in that review ($C^\top$ denotes the transpose of C). Consider the following example calculation for Y.

$$Y = MC^\top = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} [150 \quad 100 \quad 50]^\top = \begin{bmatrix} 0 \\ 50 \\ 0 \\ 0 \\ 200 \end{bmatrix}$$

In the final step, each review is ranked from the highest smoke term score to the lowest smoke term score, where higher smoke term scores imply prevalent safety defect-related language.

Rank	***j***	**This review's smoke term score**	**Review text**
1	5	200	Caught on fire instantly. Fire almost burned my house down.
2	2	50	A sharp piece broke off the side, which poses a hazard.
Tied-3	1	0	This product was excellent! Highly recommend.
	3	0	Didn't work at all! Can't believe they still sell these.
	4	0	Worked OK, but nothing special.

Researchers have debated whether an additional step should be taken to normalize the smoke term scores for each review by the word count of each review [5, 11], as this step could prevent especially long reviews from receiving excessive emphasis in the top *N*-ranked reviews. In our study, we found that this step of normalization did not meaningfully change our results. As M was a sparse matrix, longer reviews did not have substantially higher smoke term scores.

Chapter 2: Delivering business value through rapid identification of innovation opportunities in online reviews

Abstract

In recent years, online reviews have offered a rich new medium for consumers to express their opinions and feedback. Product designers frequently aim to consider consumer preferences in their work, but many firms are unsure of how best to harness this online feedback given that textual data is both unstructured and voluminous. In this study, we use text analytic tools to propose a method for rapid prioritization of online reviews, differentiating the reviews pertaining to innovation opportunities that are most useful for firms. We draw from the innovation and entrepreneurship literature and provide an empirical basis for the widely-accepted attribute mapping framework, which delineates between desirable product attributes that firms may want to capitalize upon and undesirable attributes that they may need to remedy. Based on a large sample of reviews in the countertop appliances industry, we demonstrate the performance of our technique, which performs statistically significantly better than existing methods. We validate the usefulness of our technique by asking senior managers at a large manufacturing firm to rate a selection of online reviews, and we show that the selected attribute types are more useful than alternative reviews. Our results offer insight in how firms may use online reviews to harness vital consumer feedback.

Keywords: online reviews; text analytics; data mining; innovation; business intelligence.

1.0 Introduction

This paper explores how firms can capitalize upon feedback from online reviews to derive vital insights that assist with product innovation. Manufacturers are constantly looking for ways to improve upon existing product lines or to branch into related products that can improve their competitive position. Efforts to innovate existing product lines are often nonlinear in the sense that new ideas may suddenly spur on unforeseen changes and improvements [43]. However, determining the most effective new ideas on which to build and ensuring that these ideas align with consumer demand proves difficult and complex. Ideas for product innovation may come from many sources, including brainstorming, competitive monitoring, focus groups, warranty claims, and online media [40]. MacMillan and McGrath [31, 32] propose the widely-accepted attribute mapping framework for interpreting and prioritizing innovation opportunities. The framework distinguishes between positive, negative, and neutral attributes based on simple consumer preference and then between basic, discriminator, and energizer attributes based upon consumers' likelihood to make purchasing decisions due to that sentiment. Although this framework has provided managers with a useful means of differentiating potential product attributes for decades, research has generally taken the form of case studies rather than large-scale statistical analyses [4, 35, 45].

In the years since the advent of the attribute mapping framework, firms have been faced with a new challenge in taking advantage of online media platforms. Online word-of-mouth has expanded enormously, and thousands of posts on social media and in the form of online reviews each day offer new feedback on consumers' product preferences. There is evidence that more positive online reviews are associated with greater sales [11], as 91% of consumers read online reviews to improve their understanding of products and make purchasing decisions [8]. Amidst

this avalanche of feedback, researchers and practitioners alike have sought to understand how to extract the most useful information [26]. Unfortunately, the incredible volume of online feedback makes manual review of each post unreasonable [1]; however, harnessing the incredible volume of online feedback to make the most critical of the requested product innovations could offer firms fantastic competitive advantages.

In this paper, we use the attribute mapping framework to drive rapid and automated prioritization of online reviews. We make use of a large dataset of manually coded Amazon.com reviews [34] and perform a series of text analytic methods to differentiate those reviews containing vital innovation feedback. We adapt the heuristic text analytic method proposed in Goldberg and Abrahams [18] for the detection of safety hazards in online reviews, and in this study, we apply the technique to detection of innovation opportunities. We use this method to develop lists of terms that identify reviews of interest to innovation, such as complaints or complements about key product attributes and requests for new product features. We compare our techniques to existing state-of-the-art text analytic methods, such as sentiment analysis [23, 38] and Latent Dirichlet Allocation (LDA) [6], and we find that our method greatly outperforms these techniques. The usefulness of the insights in the online reviews retrieved are verified by an assessment from senior-level managers at a large Fortune 1000-listed manufacturer of countertop appliances earning over $500 million in revenue per year.

This paper makes key contributions to several different research streams. First, this paper contributes to the theoretical stream of research on the attribute mapping framework [31, 32] by providing empirical support for the theory's application to practice. As a final stage of analysis, we asked senior-level managers at the collaborating countertop appliance manufacturer to rate the usefulness of the reviews identified by our technique versus a random chance baseline. Each

category of reviews was rated as significantly more useful than alternative online reviews. Our results show not only that the attribute mapping framework can be applied to consistently select different categories of online reviews, but also that those online reviews provide meaningful insights that assist product designers and managers in their innovation efforts. Second, this paper provides the first method for harnessing the vast volume of online reviews to provide rapid business intelligence targeted at innovation opportunities. Prior research has examined the extraction of basic product attributes from online reviews [26], but this paper is the first to bridge the gap between the discussion of attributes in these online discussions and categorizing the specific feedback for product designers to respond to the most pressing innovation opportunities (compliment, feature request, or irritator). The immense volume of online reviews presents a difficult challenge for modern firms to navigate [1]; automated tools that cut through the intimidating volume of online content and prioritize the important insights save time and focus innovation on the most important areas. Third, our study provides actionable insights for the countertop appliances industry. We present a series of words and phrases that distinguish reviews of interest to this industry that firms can implement with immediate effect to prioritize content.

The remainder of this paper is structured as follows. In our literature review, we explain the theoretical foundations for our study from MacMillan and McGrath [31, 32], and we contextualize their studies in the innovation and entrepreneurship literature. We also discuss recent literature on online reviews and media as well as the challenges associated with extracting information from these formats. We describe the key research questions that we seek to answer in this work as well as its contributions. Then, we detail our methodology, including the dataset employed in this work, the algorithms used to prioritize the most important online reviews, and the competing techniques to which we will compare our findings. We detail the results generated

from our technique and competing techniques. We validate the usefulness of our findings by a comparison performed by senior-level managers at the collaborating countertop appliances manufacturer. Finally, we conclude our paper, noting its implications as well as the potential for future work.

2.0 Literature review

A key tenant in entrepreneurship is improving offerings within the firm's product line to differentiate it from competitors and to offer a competitive advantage. Much of the entrepreneurship literature confirms this notion by arguing for an association between innovation and the success or profitability of a firm's entrepreneurial ventures [30, 42]. There is also empirical evidence that small and medium firms can improve their profitability by adapting to changes in their business environment and innovating faster than their competition, for whom these reactions may be more difficult [42, 49]. Innovation pressures have increased in recent years, as product and business model life cycles have shortened, increasing the need for firms to adapt quickly to stimuli in order to stay competitive [39, 48]. Firms that do not source inspiration for new ideas quickly and capitalize upon profitable solutions may quickly be left behind. The literature acknowledges a distinction between two types of firms: firms that generate new solutions and technologies internally and firms that adopt or build upon those solutions or technologies advanced by other firms [12]. The literature uses several different monikers to characterize this distinction, such as innovation-generating versus innovation-adopting [12], innovators versus imitators [10], or first movers versus second movers [39]. However, Pérez-Luño, Wiklund and Cabrera [39] acknowledge that these different types of firms present more of a continuum than a dichotomy; in practice, most firms initiate some original ideas internally while also monitoring their competitors to ensure that their products or services do not lack the expected attributes implemented by the competition.

The literature makes a key distinction between novelty and innovativeness [27]. Jackson and Messick [22] describe "novelty" as the extent to which a concept or instantiation differs from convention. Building upon that definition of novelty, Sethi, Smith and Park [44] describe

"innovativeness" as encapsulating whether an idea is "different from competing alternatives in a way that is valued by customers." A key distinction between these definitions is that usefulness is not a necessary component of novelty; an idea can be novel in the sense that it differs from convention without being viewed as desirable by the consumers whose needs it is ultimately meant to satisfy in application. Therefore, differentiation of potential novel ideas is vital to successful firms. Ideas that are novel without being innovative may consume resources to a greater extent than they provide revenues [16], but innovative ideas can have a transformative impact that propels firms ahead of the competition [29]. Therefore, it is of vital importance for firms to organize and prioritize their ideas so that they can focus on the strategies that provide the best fit with consumer preference.

2.1. Theoretical underpinnings

MacMillan and McGrath [31, 32] provide a model known as the attribute mapping framework for helping firms prioritize and differentiate product attributes. The attribute mapping framework has been widely applied to a variety of domains, generally in the form of case studies, such as applications in the pharmaceutical industry [35], e-business [4], and family firms [45]. The framework delineates between two dimensions of product attributes. On the horizontal axis, the framework lists positive, negative, and neutral attributes, where positive attributes are desirable to consumers, negative attributes are undesirable to consumers, and neutral attributes do not affect consumer purchasing decisions. Importantly, however, not all attributes within a row are equally influential. The vertical axis delineates between basic attributes, discriminator attributes, and energizer attributes. Basic attributes are important to consumers but unlikely to be a source of product innovation; regardless of the sentiment that consumers have concerning a

specific attribute, “basic” status implies that these attributes reflect fundamental expectations of consumers. For example, consumers may view a computer’s ability to play videos as a positive attribute; however, this feature is now so ubiquitous that virtually all consumers expect video playback as a standard feature of any new computer. Discriminator attributes may be a source of innovation; they imply that some specific group of consumers may view the given attribute as reason to choose one product over a competitive product, although this differentiation may not apply to all consumers. Product color is an example of a negative differentiator (dissatisfier); some consumers may choose a competing clothing brand if they cannot find a shirt in their favorite color. However, not all consumers will necessarily share that preference, so the dissatisfaction is only relevant among some subset of consumers. Finally, energizers are vital attributes that have near-universal reactions from consumers and create great vigor and motivation within a consumer base to make a certain purchasing decision. Revolutionary technologies may frequently fall into the category of positive energizers (exciters); for example, Apple’s original iPhone, which unified the mobile phone, media consumption, and touch display technology in a single device, offered consumers such an exciting new alternative to contemporary mobile phones that it invigorated and inspired many users to purchase iPhones. Table 1 displays the attribute mapping framework in tabular format, adapted from MacMillan and McGrath [31, 32].

	Basic	**Discriminators**	**Energizers**
Positive	Non-negotiables *Performs at least as well as competition*	Differentiators *Performs better than competition if attribute is salient to target customers*	Exciters *Performs better than competition*
Negative	Tolerables *Performs no worse than competition*	Dissatisfiers *Performs worse than competition if the attribute is salient to target customers*	Enragers *Performs worse than competition; must be corrected at any cost*
Neutral	So whats? *Does not affect purchasing decision in a meaningful way*	Parallel differentiators *Influences segment attitudes but is not directly related to performance*	N/A

Table 1. Attribute mapping framework. Adapted from MacMillan and McGrath [31, 32].

The attribute mapping framework has several useful applications for managing business innovation. The first such application is building a basic profile of a product's strengths and weaknesses with a particular understanding of the attributes of the product that are most important to consumers. Product designers and managers generally build an intuitive understanding of the strengths and weaknesses of their own products, but it is often difficult to separate their personal feelings and evaluate these products without substantial bias, necessitating the need for consumer-driven innovation [5, 13]. Often, sourcing information on product attributes from outside a product development team reveals that consumers evaluate those products differently than the internal team, and understanding these preferences better allows firms to be responsive to demand [36]. Relatedly, this step allows firms to verify their assumptions about a product's desirability. For example, comparing the attribute maps of multiple products in the same category can allow firms the ability to understand why a consumer might choose one product over another. Clarifying the reasoning for purchasing decisions

maximizes a firm's ability to adapt and respond to these preferences. Perhaps the attribute mapping framework's most valuable feature is that it provides a means of differentiating between and prioritizing different innovation opportunities. Neutral product attributes are not a promising area for major investments in product development in that consumers do not feel strongly about them in a positive or a negative sense. Altering or improving neutral attributes does not improve a product's position substantially relative to the competition because purchasing decisions are unaffected. Positive and negative attributes may or may not be important areas of emphasis; basic positive or negative attributes may be appreciated or dissatisfying to consumers respectively, but as they represent standard expectations for the product, they are not a substantial way in which to sway purchasing decisions. Instead, the most vital parts of the attribute map for prioritizing product development efforts are positive or negative discriminators or energizers, as these attributes result in different purchasing decisions. MacMillan and McGrath [31, 32] suggest that firms ought to emphasize positive energizers (exciters) and rectify negative energizers (enragers) first, as these attributes almost universally affect purchasing decisions. Positive and negative discriminators also require attention, particularly when they pertain to a large subset of consumers.

2.2. Online reviews and media

Given the proliferation of the Internet in recent years as a vibrant form of interpersonal communication, consumers look online more and more for an understanding of the products that they may be interested in purchasing. The exchange of information by users concerning their sentiments and experiences with products is referred to as word-of-mouth (WOM). The literature suggests not only that consumers read WOM frequently as a means of gathering information

about a product of interest, but also that they frequently make purchasing decisions based upon WOM [8, 11]. There are many potential venues for WOM, including social media sites (Facebook, Twitter, etc.), online review sites (Amazon, Target, Walmart, etc.), various online forums, and more. Types of WOM may vary by source; for example, Facebook posts are rather free-formed and may or may not present specific product feedback, whereas online reviews are more targeted to specific products and manufacturers. The growth of online reviews has been explosive in recent years, as the world's largest online review platforms now contain hundreds of millions of reviews [34].

Given the enormous volume of product feedback conveyed through online reviews, they present a compelling new source of information for firms. Product designers often consider many data sources when revising their products, including focus groups, consumer complaints and warranty claims, and their own ingenuity [40]. While not all consumer suggestions are feasible, experimental evidence suggests that product design that considers consumer feedback out-sells product design performed in laboratory [36]. In fact, various academic outlets have called for methods that emphasize consumer-driven innovation [5, 13]. While many practitioners surely make use of online reviews in their product development processes, online reviews are so voluminous that it is nearly impossible for practitioners to systematically read them all [2]. Therefore, methods of analyzing the content of online reviews in an automated fashion offer the possibility of enormous time savings for many firms [1, 18], and they ensure that firms are able to actually process and respond to the most critical online feedback.

Text analytics has become a popular emerging field for researchers and practitioners alike to extract meaning from online data sources. Analyzing textual data poses some unique difficulties in the sense that it does not conform to the differentiable fields of tabular-formatted

data; instead, each record is lumped into a text field. Further, many aspects of the English language (and other languages) are idiosyncratic, representing a further challenge for algorithmic approaches. Many sentiment approaches operate based on a "dictionary" of terms trained with a machine learning model that distinguishes positive text from negative text [23, 38]. These dictionaries are used to rate unseen text on a sentiment scale, where positive values such as +5 typically denote positive emotive content and negative values such as -5 typically denote negative emotive content. Applications of sentiment techniques have also been broad, such as distinguishing between positive and negative consumer attitudes [17, 46] or predicting the stock market's response to sentiment on social media platforms [7]. Text analytic methods have generally been effective at distinguishing between positive and negative consumer sentiment in online media [17, 46].

Despite its obvious appeal for analyzing online reviews and other types of online media, researchers have been quick to note some of the clear limitations of sentiment analysis [3, 18, 25]. First, because sentiment techniques tend to rely on pre-built dictionaries, they are often ineffective at coping with some of the linguistic exceptions to common conventions in the English language. For example, sentiment dictionaries typically label the word "awful" as invoking strongly negative sentiment. However, online text may use phrases such as "the price was awful good," which instead actually invokes strongly positive emotive content. Second, sentiment techniques are rather blunt tools in that positive or negative sentiment does not necessarily imply rich or interpretable content. For example, the sentence "the product was terrible" would correctly be identified by most sentiment techniques as expressing negative sentiment. However, the sentence is nonspecific and does not offer any meaningful feedback that the firm could use for remediation efforts. This problem may be compounded by domain-specific

language, such as an appliance leaking. Non-emotive or neutral terms like “leaking” offer the necessary specificity to be actionable for firms, as opposed to more emotive terms like “terrible” that do not specify a problem. Therefore, despite the appeal and success of sentiment analyses for some applications, they possess some notable limitations and deficiencies due to which more nuanced techniques may offer superior performance.

2.3 Framework for innovation opportunity discovery

We make several adaptations to the attribute mapping framework for use in the study of online reviews. The first and most fundamental change is to combine some of the initial attribute types. Using text analytic techniques, we hope to identify key words and phrases that distinguish reviews that are discussing attributes of interest to firms’ product innovation. However, the literature acknowledges that product feedback in the form of online reviews is not without bias; in particular, self-selection bias affects some of the content in online reviews [20, 28]. That is, as opposed to feedback being solicited from a random and representative sample of consumers, an online review platform is a forum in which the consumers elect to participate outside of the firm’s direct control. This facet of online reviews has both advantages and disadvantages. One advantage of this type of feedback is that consumers that self-select are likely to be highly motivated to share their opinion about a product [20] and as such will tend to offer more detailed and motivated product feedback than the average consumer. Whichever attributes of a product an online reviewer discusses in their review are likely attributes of great importance in their view. A necessary consequence is that this context makes distinguishing between discriminators and energizers quite difficult. Recall that discriminators may cause consumers to choose one product over another only if the relevant attribute is salient to them; in contrast, energizers represent

more universal differentiation and motivation. As online reviewers self-select, they are inherently consumers to whom the attributes they discuss are salient, and thus within a review it is difficult to differentiate between a discriminator that only applies to a subset of consumers and an energizer that applies more broadly. One method for delineating discriminators from energizers post-hoc may involve counting the instances of a given attribute's mention; attributes that are mentioned most often are more likely to have broad interest (energizers), while attributes that are mentioned less often may only have narrower interest (discriminators). For the purposes of the analysis of a single review, however, we combine discriminators with energizers, as a single consumer's feedback does not provide enough information to distinguish between the two. We refer to positive discriminators/energizers as "compliments" and to negative discriminators/energizers as "irritators."

A further adaptation upon the attribute mapping framework for use in analyzing online reviews is the addition of feature requests. The literature on mobile apps acknowledges that mobile app developers respond to several different types of feedback in online reviews, namely bug reports in which the developers are asked to fix a specific problem and feature requests in which the developers are asked to add a specific function [21, 41]. Feature requests represent a unique extension of the attribute mapping framework. Applied to the wider ecosystem of products, feature requests represent instances in which a consumer proposes a new feature or addition be added to a product in a future iteration. Feature requests differ from compliments and irritators, both of which refer to preexisting attributes of products, because feature requests refer to potential attributes of products that are not yet implemented. Feature requests may be either positive or negative: a consumer may request a new feature that rectifies an existing problem with the product or a feature that adds a desirable new dimension to the product.

<table>
<tr><th></th><th>Basic</th><th>Discriminators</th><th>Energizers</th></tr>
<tr><td rowspan="2">Positive</td><td rowspan="2">Non-negotiables
Performs at least as well as competition</td><td colspan="2">Compliments
Performs better than competition and is salient to target customers</td></tr>
<tr><td rowspan="2">Negative</td><td rowspan="2">Tolerables
Performs no worse than competition</td><td colspan="2">Feature requests
Potential to improve performance relative to competition</td></tr>
<tr><td colspan="2">Irritators
Performs worse than competition and is salient to target customers; must be corrected at any cost</td></tr>
<tr><td>Neutral</td><td>So whats?
Does not affect purchasing decision in a meaningful way</td><td>Parallel differentiators
Influences segment attitudes but is not directly related to performance</td><td>N/A</td></tr>
</table>

Table 2. Revised attribute mapping framework for interpreting online reviews. Adapted from MacMillan and McGrath [31, 32].

MacMillan and McGrath [31, 32] note the use of the attribute mapping framework to evaluate new ideas for potential attributes, but methods for sourcing these innovation ideas are often expensive, depending upon focus groups or consumer surveys. Exploring feature requests in online media offers product development teams actionable information based on the real-time opinions of thousands of consumers. In Table 2, we display our adapted attribute mapping framework, which is updated to reflect compliments, feature requests, and irritators. As these new types of attributes are the most striking for firms to rectify, improve upon, or otherwise consider, we focus on these core types of attributes in our text analytics study.

3.0 Research questions and contribution

In this paper, we seek to address three key research questions. First, how prevalent are compliment, feature request, and irritator attributes in online reviews? Second, to what extent can text analytic techniques be employed to extract or prioritize these attributes in online reviews, and which text analytic techniques perform best? Third, to what extent are the insights derived from these attributes useful to product innovators in practice?

We make three key contributions in this study. First, this study proposes an empirical validation of MacMillan and McGrath [31, 32]'s attribute mapping framework, which offers a theoretical view of product innovation opportunities, but evidence of the framework's effectiveness has been anecdotal or in the form of case studies rather than a large-scale empirical validation [4, 35, 45]. We examine online reviews and characterize their postings with respect to the prevalence of these important attribute types. Our findings suggest the viability of online reviews as a potential source for innovation ideas to supplement existing approaches, such as internal brainstorming or focus groups [40]. The proposed study is a novel use of state-of-the-art text mining methodology in this seldom-explored arena and a vital empirical validation of a long-standing theory in the literature. Second, this study is the first to leverage online reviews for automated extraction of innovation opportunities. Perhaps some of the most similar work was performed by Lee and Bradlow [26], who develop a text mining model for extracting marketing data from online reviews. The authors' model focuses on which product attributes are being discussed in online reviews, but it does not extend to the level of identifying the type of feedback (e.g., feature request or compliment). Additionally, the existing model is aggregated, focusing on general trends across many reviews rather than determining which individual reviews may be most interesting. The authors exhort researchers to specifically pursue data mining on "user

needs” in online reviews, a call which has not yet been filled. This study will emphasize prioritization of user needs: which reviews are the most pertinent to the feedback attribute type of interest (irritators, feature requests, or compliments). Using these techniques, practitioners can narrow the process of soliciting feedback from online reviews down to a smaller sub-sample of only the most interesting information. Third and finally, this paper presents actionable insights with immediate industry application for the countertop appliances industry. The techniques discussed in the paper and the distinctive terms generated to detect product attributes can be applied with immediate effect in this industry for the purpose of monitoring online reviews.

4.0 Methodology

4.1. Dataset and data coding

We chose Amazon.com, the world's largest e-commerce platform and largest online review platform, as the data source for this paper [34]. In collaboration with a large Fortune 1000-listed manufacturer of countertop appliances earning over $500 million in annual revenue, we chose the Amazon product categories pertaining to that firm's key product offerings and collected 733,411 total reviews pertaining to those categories. In addition to each review's basic textual content, our dataset also included the product that each review referred to, its date, its title, and its star rating on a scale of 1 to 5, inclusive. For the first stage of tagging (coding), we sought to label whether each of the reviews referred to each of the attribute types on the adapted attributed mapping framework discussed previously. We randomly chose 25,000 reviews out of this large subset for analysis. In the following, we created a scheme for delineating between online reviews that referred to feature requests, irritators, or compliments:

1) **Feature request**: The consumer is explicitly asking that the manufacturer add a specific new feature to the product or a specific modification of the product to improve the product.

Example: "I like this blender but I still need something to hold it and secure it on the base. Sometimes I feel that the top will fall off and I will be in a big trouble..."

2) **Irritator**: The consumer is unhappy or dissatisfied with a specific aspect of the product. Irritators are anything specific that the consumer dislikes or specific things that make the consumer irritated, dissatisfied, enraged, terrified, or disgusted.

Example: "ONE BIG FLAW: handle is poorly designed and has sharp ridges, so its very uncomfortable, especially when full of liquids and heavy."

3) **Compliment**: Consumer expresses happiness or satisfaction about specific aspects of the product. Compliments are specific aspects of the product that the consumer is joyful or excited about or that differentiate the product from competitor products.

Example: “I am so glad that this product came with such a nice motor... Smooth blending ability at a fraction of the cost of the more trendy blenders.”

The 25,000 randomly selected reviews were randomly distributed and presented to undergraduate business students at a major public research university for tagging. Each student was asked to tag a maximum of 200 reviews for this phase of the project to avoid a loss in data quality due to tiredness or overworking. As each review was presented to the students in an online interface, students were asked to indicate next to each review whether it referred to a given attribute type of study. To avoid the potential cognitive overload of forcing taggers to search for multiple attribute types at the same time, we staggered the tagging for each attribute type of study. That is, taggers were initially asked to identify only whether or not a review referred to a feature request; later, taggers were asked to identify only whether or not a review referred to an irritator. The separation of the tagging into binary attributes simplifies the tagging task and ideally improves the reliability of the tagging product. As the same 25,000 reviews were tagged multiple times, it is possible for one review to be labeled as referring to multiple attribute types. For example, a review could reflect both an irritator and a feature request: “the blender pitcher was awful for real cooking because it gets stuck and has to be reset. They should make a mode that periodically unsticks the blades.”

Project	Percentage of student overlap unanimous	Percentage agreement with authorities	Cohen's *K*	Fleiss, Levin and Paik [15] agreement rating	Landis and Koch [24] agreement rating
Compliments	88.0% (1,617 / 1,837)	89.3% (125 / 140)	0.786	Excellent agreement	Substantial agreement
Feature requests	91.1% (8,890 / 9,759)	93.0% (186 / 200)	0.860	Excellent agreement	Almost perfect agreement
Irritators	85.8% (5,104 / 5,950)	86.2% (181 / 210)	0.724	Fair to good agreement	Substantial agreement

Table 3. Tagger reliability descriptive statistics.

An important characteristic of the above tagging scheme is that each feature request, irritator, or compliment must be specific and explicit. Many consumer reviews may reveal some generic sentiment that a consumer has regarding a certain product, but nonspecific sentiment is not an actionable component of the attribute mapping framework. For example, some reviews might nonspecifically state that "this blender was awful, definitely don't buy it." Although the reviewer clearly complains about the quality of the product, as they do not express a specific issue with the product beyond their general sentiment, the review is of little use to manufacturers. Taggers were instead encouraged to focus only on specific feedback from consumers, which is more useful for innovation purposes.

The data tagging process for this project was examined on several levels to ensure the reliability of the output. First, as reviews were assigned to taggers at random, some reviews are tagged multiple times. In these instances, it is necessary to reconcile multiple tags for one review, and particularly in the case of conflicting tags, to render a final decision. Following the example of prior work [18, 50], we follow a "majority conservative" decision rule, in which we choose the majority vote in the event of any disagreement. However, if the votes are tied, then

we choose the target classification (i.e., compliment, feature request, or irritator). This strategy avoids false negatives when there is any doubt as the correct code for any review. The overlapping of taggers provides us an opportunity to assess reliability because we can compute the percentage of overlapping tags on which taggers agreed with one another. A higher percentage of the overlapping tags on which the taggers were unanimous implies more reliable tagging. Second, for external validation, we asked for tagging support from senior-level managers at the collaborating countertop appliances manufacturer. These expert taggers serve as authorities to validate the tagging of our students. We recruited six senior-level managers to provide tags on each of the three projects (compliments, feature requests, and irritators), assigning two managers to each project, and we provided each with the same tagging protocol as the students. Each authority tagger was supplied with a subset of reviews that students had tagged; the subset was selected using a stratified random sample such that each binary classification was equally likely. In addition to each binary classification, we also asked the managers to rate how useful they felt the review would be to their innovation process on a three-point scale [44]: (1) not useful at all; (2) a bit useful; and (3) very useful. In doing so, we have a mechanism by which to assess the real-world applicability of these reviews to innovation processes. We can compute two statistics to measure the agreement between student taggers and authority taggers: first, the percentage of tags on which the student taggers and the authority taggers agreed, and second, Cohen's *K*, which compares the agreement percentage to random chance agreement. A higher agreement percentage and, relatedly, a Cohen's *K* value closer to 1 indicate more reliable tagging. In Table 3, we present these statistics across the three tagging projects. Each project scored high levels of agreement on all measures, suggesting that the taggers consistently identified the same types of reviews in each attribute.

4.2. Term curation

In the text analytics literature, Abrahams et al. (2015) and, later, Goldberg and Abrahams (2018) describe a supervised learning approach for curating "smoke terms" that delineate reviews that mention product safety hazards from reviews that do not. "Smoke terms" refer to distinctive words and phrases that occur especially prevalently in reviews that contain these concerns. This technique has been applied with great success for detecting product safety hazards across several industries, including baby cribs [37], dishwashers [25], and joint and muscle pain treatments [3]. As the lexical properties of online reviews in each industry tend to differ, smoke term lists are typically tailored to a specific domain. For example, the term "choking hazard" may be a distinctive issue in baby cribs industry [37], but it does not likely generalize to dishwasher reviews.

In this paper, we adapt the aforementioned methodology for use in detection of product innovation opportunities in online reviews. A schematic of this methodology is displayed in Figure 1. We use the technique multiple times to generate lists of applicable terms for each attribute type of interest (compliments, feature requests, and irritators). We initially separate our dataset into three approximately equally sized portions. The first portion (A) is the training set, which is used to generate a basic linguistic understanding of the online reviews and determine candidate terms to distinguish features of interest [14, 18]. The second portion (B) is the curation set, which is used to test the candidate smoke terms and choose those that provide the highest level of performance. The third portion (C) is the validation set, which is used to provide measures of the efficacy of the term lists.

After our dataset is divided into three portions, we use information retrieval techniques to identify candidate smoke terms that may distinguish reviews of interest to a given attribute type.

Each review is classified into two binary categories; for example, “feature request” versus “no feature request” or “irritator” versus “no irritator.” Previous works using the smoke term methodology suggest that the CC score proposed by Fan, Gordon and Pathak [14] performs well for generating initial candidate smoke terms [1, 18]. The CC score algorithm assesses the “relevance” of terms based on how often they occur in relevant versus irrelevant reviews; for example, a term that occurs frequently in “feature request” reviews and is infrequent otherwise would receive a high score [14]. We use the CC score algorithm to generate relevance scores pertaining to all unigrams (one-word), bigrams (two-words), and trigrams (three-words) in our training set (1). These relevance scores are retained, as they serve as weights in a later stage.

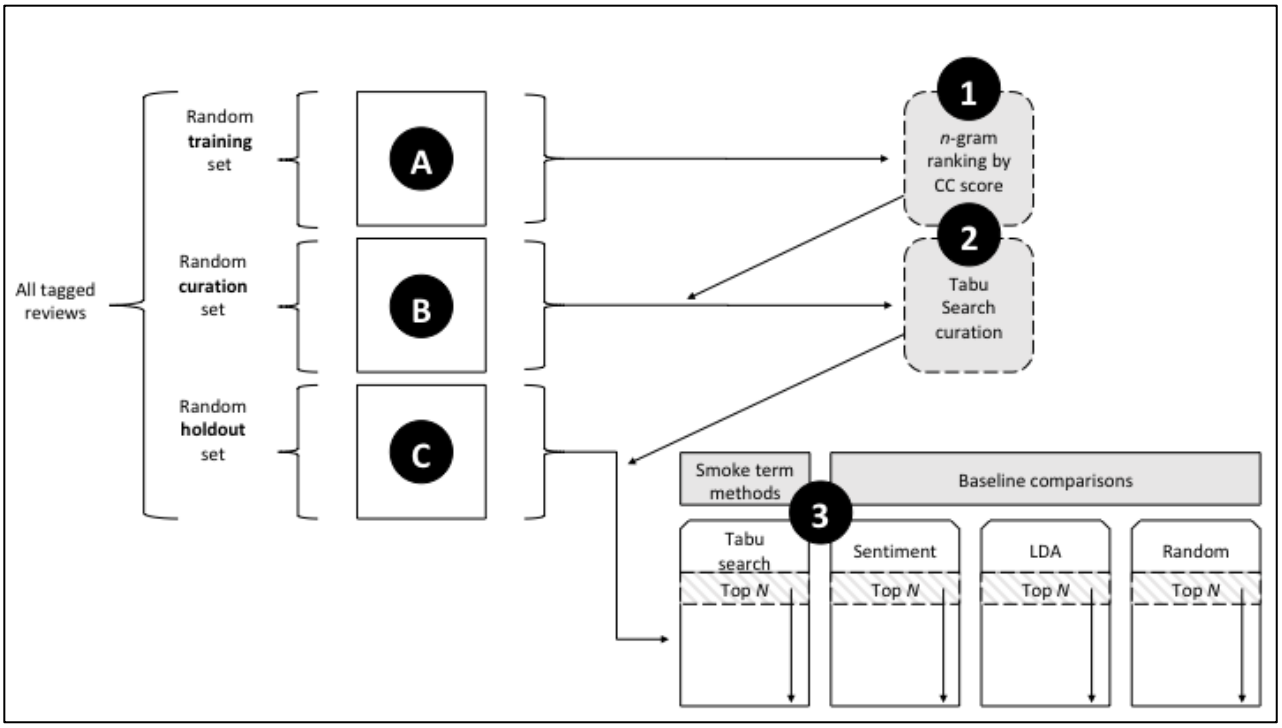

Figure 1. Schematic of text analytic methodology.

Next, we make use of the curation set to refine the initial set of candidate smoke terms (2). For this step, we use the Tabu search heuristic proposed in Goldberg and Abrahams [18].

The authors describe a method by which the algorithm is configured to maximize the precision of potential smoke term lists based on the curation set. That is, if all reviews in the curation set are ranked by a score of the number of instances of each smoke term in a given review multiplied by the relative terms' relevance scores (weights), then precision measures the number of reviews in the top *N*-ranked set that refer to true instances. The Tabu search algorithm tests multiple combinations of terms iteratively, each time calculating the precision obtained in the curation set. The algorithm adds terms to the list that improve precision and removes those that harm precision. As the goal of this technique is to choose the top-ranking reviews to read, Goldberg and Abrahams [18] suggest optimizing precision in the top 100-ranked or 200-ranked reviews or using an average of those precision values. Following their example, we maximize the average level of precision in the top 100-ranked and 200-ranked reviews.

The final step in the procedure involves the incorporation of the holdout set, which is unseen by the algorithm thus far. Using the terms generated by the Tabu search heuristic, we test how the selected terms perform given this unseen data using the aforementioned scoring procedure the rank the reviews from most likely to least likely to be of interest to innovation processes (3). The step of testing the efficacy of the terms on unseen data helps to ensure that the terms are not "overfit" and that they generalize well to new datasets [1, 18].

4.3 Competing text analytic techniques

To ensure the efficacy of our results using the aforementioned text analytic techniques, we take the further step of comparing our methods to several alternative text analytic methods. The first such method that we consider is sentiment analysis, which is used to assess the emotive content in text. Sentiment analysis is widely applicable in many domains, such as predicting the

results of political elections [9] or detecting safety hazard discussions in online media [1]. We use two existing sentiment techniques, AFINN [38] and the Harvard General Inquirer [23] to rank the online reviews in our dataset by their emotive content, and we compare this ranking to our technique. Each sentiment analysis technique assesses text on multiple emotive dimensions, such as positivity, negativity, strength, passivity, pleasure, pain, etc. In this case, we use positivity to assess positive sentiment (compliments) and negativity to assess negative sentiment (irritators).

Additionally, Latent Dirichlet Allocation (LDA), first proposed by Blei et al. [6], is a popular text analytic method for mining "topics" in online reviews. The theory underlying LDA suggests that each text corpus is made up of a mixture of several topics. Each document in that corpus (for example, an online review) may contain only some of those topics. Some documents may discuss only one topic, while others may discuss a blend between several topics. LDA is an unsupervised machine learning technique that uses a three-level hierarchical Bayesian model to estimate which terms comprise which topics. LDA has many applications, such as conducting analyses of online chatter about brands in online media [47] and characterizing the topics that make up consumer discussions in service industries [19, 51]. We use LDA to determine lists of topics reflecting each attribute type and the extent to which these topics are predictive of innovation opportunities.

5.0 Results

5.1. Tagging results

After reconciling the tags completed by the student and authority taggers, we first present a description of the tagging results in Table 4. Compliments, feature requests, and irritators were all well-represented in online reviews. Compliments were the best represented at 32.9% of our dataset, whereas irritators constituted 18.0% of the dataset, and feature requests constituted 5.4% of the dataset (note that these figures are not additive as a single review could be coded as referring to multiple attributes). Compliments were most prevalent in reviews that scored high star ratings; however, even 1-star and 2-star reviews were well represented, as even generally critical reviews sometimes note some positive attribute of a product. Irritators were most associated with negative reviews, although even some high-scoring reviews noted some irritations with otherwise satisfactory products. Feature requests had the weakest association with star ratings; they were slightly more prevalent in high-scoring reviews, but they were most common with 4-star reviews. We also found evidence that these proportions vary by firm: the firm that collaborated with us had far more compliments, slightly more feature requests, and slightly fewer irritators than average. Interestingly, as the firm generally received very high star ratings, most irritators were actually found in 5-star reviews. Therefore, simply analyzing reviews with extreme star ratings is a risky strategy, as each attribute type appears across the full continuum of star ratings.

	All firms			Collaborating firm		
Star rating	**Compliments**	**Feature requests**	**Irritators**	**Compliments**	**Feature requests**	**Irritators**
1	800	163	1,237	20	18	44
2	646	142	887	9	16	25
3	1,277	222	858	24	21	29
4	2,340	486	932	285	42	65
5	3,176	341	593	594	29	191
Total	8,239 (32.9% of reviews)	1,354 (5.4% of reviews)	4,507 (18.0% of reviews)	932 (42.1% of reviews)	126 (5.7% of reviews)	354 (16.0% of reviews)

Table 4. Star rating distribution for compliments, feature requests, and irritators.

Our findings suggest that online reviews are a viable medium for discovering innovation opportunities. Each attribute type identified in the adapted attribute mapping framework (see Table 3) is present in online reviews and was reliably coded. However, each attribute is present only in a minority of reviews, and given the great volume of online content, prioritizing the content with text analytic techniques to access the most relevant feedback is crucial.

5.2 Text analytics results

We employed the aforementioned heuristic text analytic method proposed by Goldberg and Abrahams [18] in order to generate unigram, bigram, and trigram terms for each attribute of interest. We display the top five unigrams, bigrams, and trigrams generated by this technique for each attribute in Table 5. Like the findings of prior work, the terms seem to reflect domain-specific jargon [1] as well as narrative structure [18]. Several of the terms, such as “lid,” “the alarm,” “the unit,” or “the top,” have particular meanings in the countertop appliances industry in that they refer to specific attributes or components of a product. Even though these terms were

all instances of feature requests or irritators, indicating that the consumer was noting some area of improvement for the product, none of these terms explicitly states a negative experience. These words and phrases are unlikely to be well recognized by sentiment approaches, which are not tuned to the specific nuances of the countertop appliances domain. In the context of this domain, however, consumers only tend to use these terms when they are describing an issue with some component of the product. An irritator might complain "I couldn't get **the top** to stay on," and this type of usage, which is largely non-emotive, is hugely prevalent in domain-specific WOM [18]. In addition, many of the terms do not refer specifically to a product attribute, but they instead identify the narrative in which the customer describes their experience. For example, the phrase "stars is because" reflects instances in which the reviewer attempts to justify the star rating in their review by noting some experience with an aspect of the product that affected their final rating. Reviewers using this phrase do so with the expectation that others will read their review, and they preemptively justify their star rating to those readers.

Panel A: Top unigram, bigram, and trigram compliment terms.

Unigram	Weight	Bigram	Weight	Trigram	Weight
easy	96,308	easy to	85,383	easy to clean	63,277
highly	36,742	very easy	39,662	easy to use	53,435
durable	19,122	so easy	26,518	i love this	26,117
fantastic	17,462	fast and	17,855	so easy to	21,900
owned	16,170	clean i	15,247	highly recommend it	12,500

Panel B: Top unigram, bigram, and trigram feature request terms.

Unigram	Weight	Bigram	Weight	Trigram	Weight
wish	33,510	wish it	28,360	i wish it	19,662
needs	14,100	wish the	19,993	have been nice	15,945
perhaps	13,878	would be	17,176	stars is because	14,598
lid	12,506	been nice	15,945	could have been	12,929
change	10,319	the alarm	13,934	if it had	12,929

Panel C: Top unigram, bigram, and trigram irritator terms.

Unigram	Weight	Bigram	Weight	Trigram	Weight
return	49,499	would not	40,174	do not buy	36,444
disappointed	38,754	does not	38,513	not buy this	31,392
first	38,406	the unit	36,455	i have to	29,546
stopped	31,745	not recommend	35,266	piece of junk	29,385
way	30,577	the top	29,473	would not recommend	28,642

Table 5. Top terms generated by the Tabu search heuristic [18].

Using the unseen holdout set, we compare our technique to several other text analytic techniques to benchmark its performance. In addition to sentiment analyses [23, 38], we also used LDA [6] to generate topics that may be predictive of the attributes from the attribute mapping framework. We ran the LDA algorithm for 1,500 iterations and generated 10 topics of

15 words each, displayed in Table 6[1]. We manually labeled each topic based on its contents [19]. Most topics identified different types of products, but we identified topic #2 as denoting negative product experiences, which we used to predict irritators, and topic #4 as denoting positive product experiences, which we used to predict compliments. None of the topics seemed to relate to feature requests.

We use each of these methods to rank the reviews in our dataset from most likely to least likely to contain a compliment, feature request, and/or irritator. Assessing the efficacy of these techniques involves choosing an arbitrary cutoff (the top *N*-ranked reviews for a given technique) and computing the number or percentage of true instances of each attribute within that cutoff. In Table 7, we show the performance of each technique at the top 200 reviews, following the example of previous works [1, 18], as this volume of content might be considered manageable for many firms. The top performing technique for each attribute is indicated in **bold**. For each of compliments, feature requests, and irritators, our domain-specific terms far outperform the competing techniques. LDA was the nearest performing alternative for compliments, but it was the worst performing alternative for irritators. Simple star ratings bested both LDA and sentiment analysis at predicting irritators, but they were very poor at predicting feature requests. For each attribute, we compared our domain-specific techniques to the competing techniques using a chi-squared test; each technique significantly differed from each competing technique at the 0.001 level. In Figure 2, Figure 3, and Figure 4, we display the performance of each technique over a range of possible cutoffs.

[1] We tested running LDA for various numbers of topics ranging from 2 to 25. When fewer topics were identified, none seemed to refer to positive or negative product experiences. When more topics were identified, LDA generated more topics that pertained to types of countertop appliances, such as food processors, microwaves, etc., or redundant topics.

Topic	Top terms
#1: Pans and cookware	pan, set, pans, stick, use, it, non, cooking, cook, cookware, great, heat, well, clean, stainless
#2: Negative product experience	one, product, amazon, back, it, unit, get, first, reviews, new, could, time, buy, made, replacement
#3: Water filters	water, air, kettle, it, filter, unit, room, use, filters, smell, much, really, dust, clean, fan
#4: Positive product experience	it, great, easy, use, love, product, one, works, well, recommend, bought, price, opener, clean, gift
#5: Purchasing narrative	one, years, it, bought, needed, old, used, new, last, year great, use, another, model, still
#6: Blenders/juicers	blender, ice, it, use, make, cream, clean, juicer, machine, great, easy, juice, blade, food, get
#7: Slow/rice cookers, mixers	rice, mixer, cooker, it, pot, use, time, one, cooking, bowl, cook, slow, great, love, used
#8: Coffee makers	coffee, cup, maker, water, it, machine, hot, pot, use, carafe grinder, one, great, brew, makes
#9: Popcorn and waffle makers	popcorn, easy, it, use, pop, great, waffle, clean, cooking, one, time, cook, make, oil, waffles
#10: Toasters	toaster, it, oven, toast, lid, top, one, use, well, unit, makes, get, time, bread, nice

Table 6. 10-topic analysis output from LDA [6].

	Number (percentage) of true instances in top 200-ranked reviews		
Technique	**Compliments**	**Feature requests**	**Irritators**
Unigrams	171 (85.5%)	62 (31.0%)	**148 (74.0%)**
Bigrams	**173 (86.5%)**	87 (43.5%)	140 (70.0%)
Trigrams	**173 (86.5%)**	**103 (51.5%)**	135 (67.5%)
LDA	116 (58.0%)	--	67 (33.5%)
AFINN	95 (47.5%)	27 (13.5%)	77 (38.5%)
Harvard GI	85 (42.5%)	31 (15.5%)	63 (31.5%)
Star ratings	102 (51.0%)	8 (4.0%)	97 (48.5%)

Table 7. Performance of each technique within the top 200-ranked reviews.

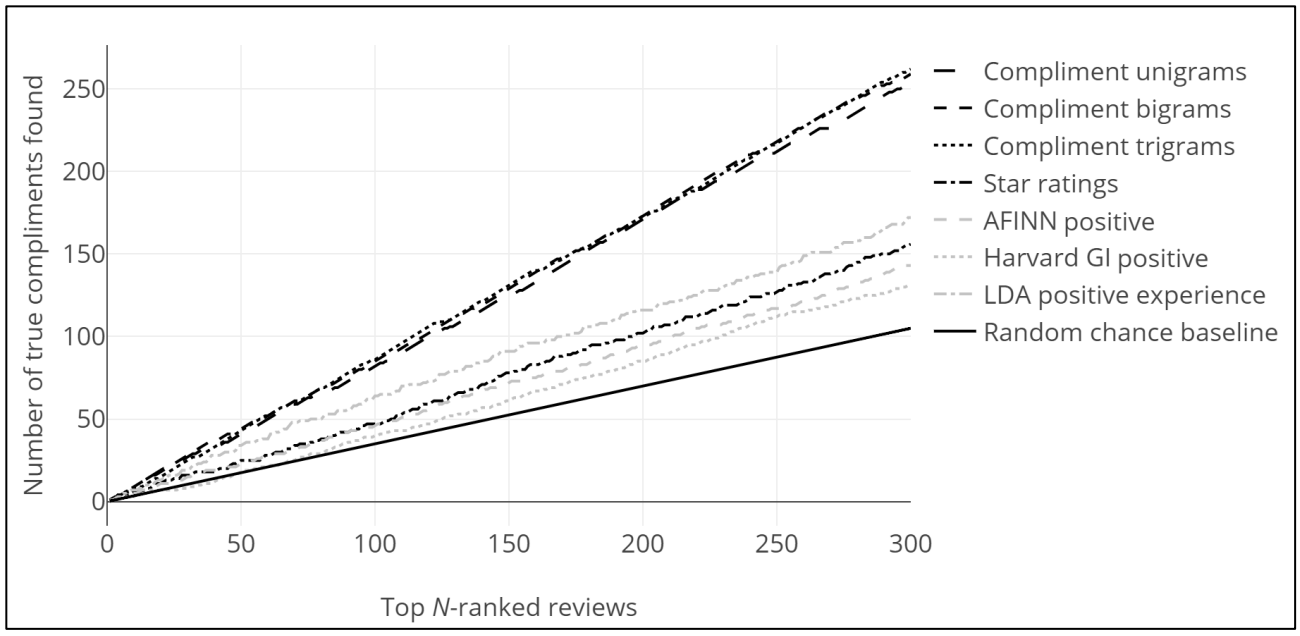

Figure 2. Lift chart of text analytic performance predicting compliments.

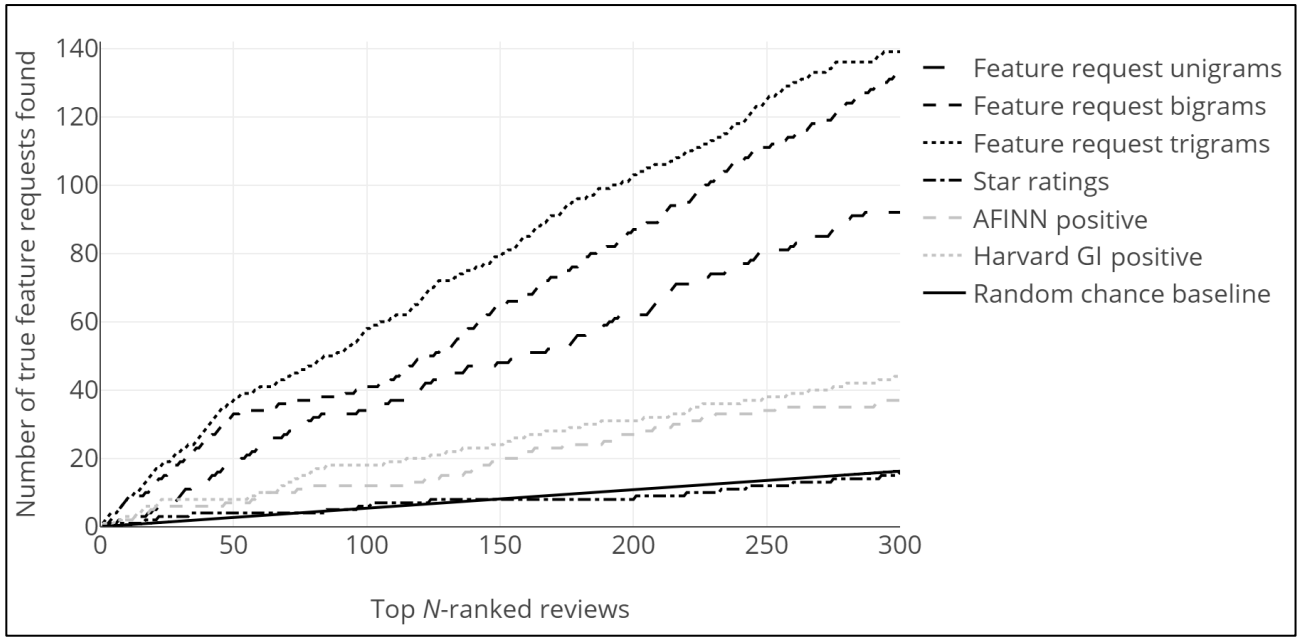

Figure 3. Lift chart of text analytic performance predicting feature requests.

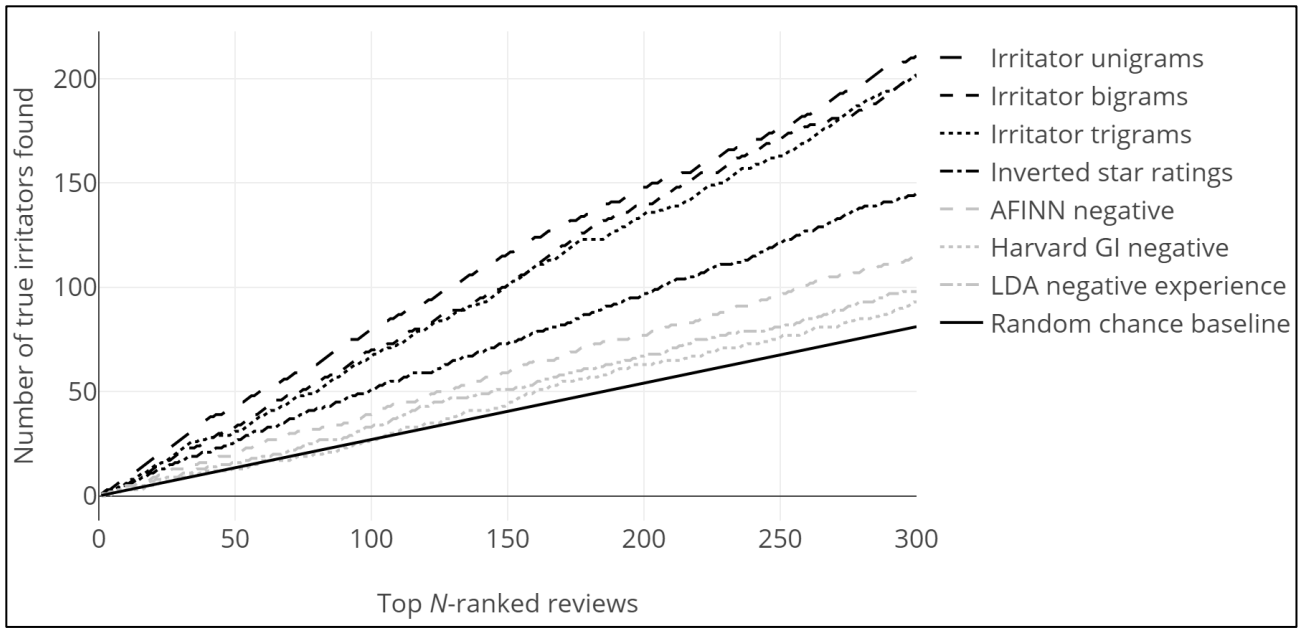

Figure 4. Lift chart of text analytic performance predicting irritators.

5.3 Case study

In the following, we show a short case study for deploying automated detection of innovation opportunities in online reviews. We chose a line of coffee makers by a competitor of the collaborating countertop appliance manufacturer for our case study. We filtered the reviews pertaining to those products, and we ranked those reviews based on the domain-specific terms generated previously. In Table 8, we show a selection of top-scoring reviews for each of these attribute types. Terms in the domain-specific dictionaries are indicated in **bold**, and specific feedback for the firm's product offerings is underlined. As previous work has suggested, the reviews tend to follow a narrative structure in which the reviewer details their experience with the product [18]. Each of top-ranking compliment reviews praised the product's ease of use, and each top-ranking feature request review requested that the product be redesigned to handle different or larger cup sizes, a concern shared by the top compliment review. The second feature

request review advises readers to purchase competing products on that basis. The top two irritator reviews both express frustration at the product's leaking, possible due to a faulty gasket, while the third irritator review complains that the brews made by their machine have shrunken over time. The first and third irritator reviews indicate that the irritation affects their purchasing decisions, as they otherwise enjoy the product but would opt for an alternative if the problem persists.

Top compliments	Top feature requests	Top irritators
Overall **I love this** machine. It looks sleek; very **easy to use**. My only complaint is the amount of water that is required to be kept in the reservoir even for a small amount of coffee.	I like my [brand name]. It makes a consistent cup of coffee. This was the cheaper version of the different [brand name] machines available and Im very glad that I got it. **I do wish there** was a way to **change** the amount of coffee brewed. There is I suppose, but that would mean a more expensive machine. It makes about half a cup of what I would consider a normal cup of coffee. Other than that I am very happy with the machine…	… overflows water and grounds into the receiving cup due to a faulty gasket above the upper needle. No instructions tell you that **you have to** lower the gasket ring above the needle each time you use the [brand name] - forget and the coffee is not fit to drink, lower it too far and it comes off. This is not an improvement. This is a great item if it worked as advertised. I **would not recommend** it at this time.
This this is amazing. I honestly cannot say one bad thing about it. Extremely **easy to use,** great coffee (of course that also depends on what kind of coffee you use), and perfect size for my desk .	This unit works great. It's simple and functional. However, if you're purchasing something for your home, opt for a better model or you will be disappointed. This is **functional** and small for an office or an area away from home but for the home, trust me, you want the model **with different** cup sizes and the water reservoir because it makes the experience so much better…	We loved our [brand name], for about 18 months. Then it started **leaking** from the bottom. We tried adjusting the gasket where the needle comes in per several suggestions. then we took it apart and tried to clean and adjust a gasket inside. Still **leaks**…
I bought this for my daughter, she loves it **easy to use** and convenient, bought some of the pods that can be used with regular coffee and saves some money	I bought this to use at work and I love it. I wanted something a little bit compact to use in my office. The only negitive is that it takes 3 minutes to get my coffee. **I do wish it** would hold a few cups of water so **it would be** more instant like the one I have at home.	like other reviews, we have found that the more we use this unit, the smaller our cups of coffee get. not sure what the issue is. i am hoping to find a fix or that the cup size will stop shrinking. if not i will **return the unit**. other than that issue, it works great and is perfect for our house, where i like tea and my wife likes coffee.

Table 8. Top-scoring reviews by attribute type for coffee makers.

These reviews present clear instances in which consumer purchasing decisions are directly related to product attributes. Using the rapid feedback from a small sampling of prioritized reviews, the firm can immediately interpret their position in the context of the revised

attribute mapping framework, shown in Table 9. The issues with brew sizes and the leaking/faulty gasket may sway many consumers, including those reading the online reviews, to purchase alternative products. The firm may prioritize these fixes in their product development process. Meanwhile, the firm can market its product as being easy to use, which makes a positive impression on their customers.

<table>
<tr><th></th><th>Basic</th><th>Discriminators</th><th>Energizers</th></tr>
<tr><td rowspan="2">Positive</td><td rowspan="2">Non-negotiables</td><td colspan="2">Compliments
Ease of use</td></tr>
<tr><td colspan="2">Feature requests
Different/larger brew sizes</td></tr>
<tr><td rowspan="2">Negative</td><td rowspan="2">Tolerables</td><td colspan="2">Irritators
Brew size decreasing over time
Leaking/faulty gasket</td></tr>
<tr><td>Neutral</td><td>So whats?</td><td>Parallel differentiators</td><td>N/A</td></tr>
</table>

Table 9. Attribute map for coffee makers based on online review intelligence.

5.4. Validation of usefulness

Authority taggers were asked to offer their opinion on the usefulness of each review that they tagged in a three-point scale. In doing so, we enable a comparison of the usefulness of each attribute type (compliments, feature requests, and irritators) versus other online reviews. Table 10 presents the authority taggers' tagging counts for each of these attribute types. In each attribute type, the authority taggers indicated that they found the target classification reviews more useful than alternative reviews that they were provided. As our data is ordinal and non-normal, we assess the difference between each attribute and alternative reviews statistically using

a Mann-Whitney U test [33]. These statistical tests indicated that tags for each attribute type, compliments, feature requests, and irritators, differed significantly from alternative reviews at the 0.001 level. This statistical evidence suggests that online reviews are a meaningful source of innovation ideas that are beneficial to product development teams, and it provides an empirical basis that the attribute mapping framework applies in this domain [31, 32]. One potential source of bias in the authority taggers' assessments of usefulness is that they tagged for attribute mapping components and usefulness at the same time. For robustness against this potential cross-contamination, we used a holdout authority tagger from the collaborating countertop appliance manufacturer for one further round of tagging. We presented this tagger with a stratified random sample of reviews in which 1/6 had been identified as compliments, 1/6 had been identified as feature requests, 1/6 have been identified as irritators, and the remaining 1/2 did not fall into any attribute type of interest. The holdout tagger only tagged for usefulness on the three-point scale without regard for the attribute mapping constructs. This tagger's results are displayed in Table 11. Like the other authority taggers, this tagger rated compliments, feature requests, and irritators as more useful than alternative reviews, and their tags in each attribute type significantly differed from the alternative reviews at the 0.001 level.

	Number (percentage) of authority tags by attribute type and usefulness					
	Compliments		Feature requests		Irritators	
Usefulness	**Compliment**	**No**	**Feature request**	**No**	**Irritator**	**No**
Not useful at all	7 (10.6%)	39 (52.7%)	36 (36.0%)	71 (71.0%)	22 (20.8%)	62 (59.6%)
A bit useful	35 (53.0%)	30 (40.5%)	46 (46.0%)	25 (25.0%)	62 (58.5%)	34 (32.7%)
Very useful	24 (36.4%)	5 (6.8%)	18 (18.0%)	4 (4.0%)	22 (20.8%)	8 (7.7%)
Total	66	74	100	100	106	104

Table 10. Authority taggers' perceptions of online review usefulness.

	Compliment	**Feature request**	**Irritator**	**None**
Not useful at all	10 (47.6%)	4 (20.0%)	4 (19.0%)	56 (82.4%)
A bit useful	8 (38.1%)	8 (40.0%)	13 (61.9%)	11 (16.1%)
Very useful	3 (14.2%)	8 (40.0%)	4 (19.0%)	1 (1.4%)
Total	21	20	21	68

Table 11. Holdout tagger's perceptions of online review usefulness.

6.0 Conclusions

This paper presents the first large-scale empirical validation of the popular attribute mapping framework by MacMillan and McGrath [31, 32]; as opposed to case studies, we find enormous statistical evidence not only that consumers post feedback that aligns with the attribute mapping framework, but also that firms can glean valuable insights from that feedback. This empirical validation lends credence to many other works that incorporate this theoretical framework [4, 35, 45]. Our text analytic results suggest that the exaptation of the Goldberg and Abrahams [18] methodology applies to product innovation opportunities in online reviews, responding to the call by Lee and Bradlow [26] for the development of data mining technologies that directly consider user needs. Particularly since hastened product and business life cycles have increased innovation pressures on firms [39], the capability to rapidly source innovation-related feedback from online media is imperative. There are a plethora of potential applications for practitioners in these rapid prioritization technologies. Most obviously, the proposed technique allows for firms to quickly and easily source innovation-related feedback from a mass of consumers and map it into a verified framework for organizing and prioritizing product attributes. In addition to their own products, firms can perform analyses of competing products to discern the source of competitive advantages and thereby to obtain a superior position. This analysis can be performed prospectively, searching for new insights using customer-driven feedback, or retrospectively, using consumer-driven feedback to verify or update an existing perception.

Our work is subject to several limitations. First, we relied on a large team of volunteer taggers to help code the data used in this study, and our supervised learning technique relied on these volunteers' opinions, which introduces a source of bias and variability. Delineating

between one attribute type and another (e.g., “feature request” versus “no feature request”) is a somewhat subjective process, as it relies upon each individual’s reading of the online review. We used several safeguards in this work to minimize this variability as much as possible, including providing the taggers with a detailed protocol document and comparing the tags to authorities, with whom they had considerable agreement. Second, our techniques should serve as a useful component of innovation efforts, but they should not be used alone. Prior work has extolled the virtues of many sources for innovation ideas, including brainstorming, focus groups, consumer surveys, warranty claims, and competitive monitoring [40]. Our technique helps firms to harness a massive volume of online consumer-driven data, but it does not serve to replace existing methods.

In future work, in addition to prioritizing online reviews using the framework described in this paper, practitioners may also like to aggregate reviews of similar topics together to understand overarching trends. After our technique is run as an initial stage, these techniques may, for example, allow firms to allocate reviews to relevant teams that can directly address innovation opportunities. This step is beyond the scope of this paper, but a topic mining technique such as LDA [6] or a simpler bag-of-words model may assist in this process. Another possible extension of this work concerns extending our framework to other forms of online media, such as social media, news, and/or forum posts. Online reviews are clearly associated with specific products, so it is clear which reviews pertain to relevant products; on the broader web, further techniques could assist in rapidly sifting through different forms of textual data and identifying pressing innovation-related content.

References

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

[2] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decision Support Systems, 54(1) (2012) 87-97.

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

[4] R. Amit, C. Zott, Value creation in e-business, Strategic Management Journal, 22(6-7) (2001) 493-520.

[5] L.A. Bettencourt, A.W. Ulwick, The customer-centered innovation map, Harvard Business Review, 86(5) (2008) 109.

[6] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning research, 3(Jan) (2003) 993-1022.

[7] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science, 2(1) (2011) 1-8.

[8] BrightLocal, Local Consumer Review Survey 2016, in, (2016).

[9] A. Ceron, L. Curini, S.M. Iacus, G. Porro, Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France, New Media & Society, 16(2) (2014) 340-358.

[10] M.-H. Chang, J.E. Harrington Jr, Innovators, imitators, and the evolving architecture of problem-solving networks, Organization Science, 18(4) (2007) 648-666.

[11] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: Online book reviews, Journal of Marketing Research, 43(3) (2006) 345-354.

[12] F. Damanpour, J.D. Wischnevsky, Research on innovation in organizations: Distinguishing innovation-generating from innovation-adopting organizations, Journal of Engineering and Technology Management, 23(4) (2006) 269-291.

[13] K.C. Desouza, Y. Awazu, S. Jha, C. Dombrowski, S. Papagari, P. Baloh, J.Y. Kim, Customer-driven innovation, Research-Technology Management, 51(3) (2008) 35-44.

[14] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decision Support Systems, 40(2) (2005) 213-233.

[15] J.L. Fleiss, B. Levin, M.C. Paik, Statistical methods for rates and proportions, (John Wiley & Sons, 2013).

[16] D.C. Galunic, S. Rodan, Resource recombinations in the firm: Knowledge structures and the potential for Schumpeterian innovation, Strategic Management Journal, (1998) 1193-1201.

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

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

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

[20] N. Hu, J. Zhang, P.A. Pavlou, Overcoming the J-shaped distribution of product reviews, Communications of the ACM, 52(10) (2009) 144-147.

[21] C. Iacob, R. Harrison, Retrieving and analyzing mobile apps feature requests from online reviews, in: Proceedings of the 10th Working Conference on Mining Software Repositories, (IEEE Press, 2013), pp. 41-44.

[22] P.W. Jackson, S. Messick, The person, the product, and the response: Conceptual problems in the assessment of creativity, Journal of Personality, 33(3) (1965) 309-329.

[23] E.F. Kelly, P.J. Stone, Computer recognition of English word senses, (North-Holland, 1975).

[24] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics, (1977) 159-174.

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

[26] T.Y. Lee, E.T. Bradlow, Automated marketing research using online customer reviews, Journal of Marketing Research, 48(5) (2011) 881-894.

[27] J.J. Li, X.-P. Chen, S. Kotha, G. Fisher, Catching fire and spreading it: A glimpse into displayed entrepreneurial passion in crowdfunding campaigns, Journal of Applied Psychology, 102(7) (2017) 1075.

[28] X. Li, L.M. Hitt, Self-selection and information role of online product reviews, Information Systems Research, 19(4) (2008) 456-474.

[29] J.C. Linder, S. Jarvenpaa, T.H. Davenport, Toward an innovation sourcing strategy, MIT Sloan Management Review, 44(4) (2003) 43-50.

[30] G.T. Lumpkin, G.G. Dess, Clarifying the entrepreneurial orientation construct and linking it to performance, Academy of Management Review, 21(1) (1996) 135-172.

[31] I.C. MacMillan, R.G. McGrath, Discover your products' hidden potential, Harvard Business Review, 74(3) (1996) 58-73.

[32] I.C. MacMillan, R.G. McGrath, Discovering new points of differentiation, Harvard Business Review, 75 (1997) 133-145.

[33] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Annals of Mathematical Statistics, (1947) 50-60.

[34] J. McAuley, R. Pandey, J. Leskovec, Inferring networks of substitutable and complementary products, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM, 2015), pp. 785-794.

[35] R.G. McGrath, A. Nerkar, Real options reasoning and a new look at the R&D investment strategies of pharmaceutical firms, Strategic Management Journal, 25(1) (2004) 1-21.

[36] D.L. Meadows, Estimate accuracy and project selection models in industrial research, Industrial Management Review, 9(3) (1968) 105.

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

[38] F.Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, arXiv preprint arXiv:1103.2903, (2011).

[39] A. Pérez-Luño, J. Wiklund, R.V. Cabrera, The dual nature of innovative activity: How entrepreneurial orientation influences innovation generation and adoption, Journal of Business Venturing, 26(5) (2011) 555-571.

[40] J. Pruitt, T. Adlin, The persona lifecycle: keeping people in mind throughout product design, (Elsevier, 2010).

[41] Z. Qiao, G.A. Wang, M. Zhou, W. Fan, The Impact of Customer Reviews on Product Innovation: Empirical Evidence in Mobile Apps, in: Analytics and Data Science, (Springer, 2018), pp. 95-110.

[42] N. Rosenbusch, J. Brinckmann, A. Bausch, Is innovation always beneficial? A meta-analysis of the relationship between innovation and performance in SMEs, Journal of Business Venturing, 26(4) (2011) 441-457.

[43] H. Sarooghi, D. Libaers, A. Burkemper, Examining the relationship between creativity and innovation: A meta-analysis of organizational, cultural, and environmental factors, Journal of Business Venturing, 30(5) (2015) 714-731.

[44] R. Sethi, D.C. Smith, C.W. Park, Cross-functional product development teams, creativity, and the innovativeness of new consumer products, Journal of Marketing Research, 38(1) (2001) 73-85.

[45] D.G. Sirmon, M.A. Hitt, Managing resources: Linking unique resources, management, and wealth creation in family firms, Entrepreneurship Theory and Practice, 27(4) (2003) 339-358.

[46] S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, Journal of Management Information Systems, 29(4) (2013) 217-248.

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

[48] A.H. Van de Ven, M.S. Poole, Explaining development and change in organizations, Academy of Management Review, 20(3) (1995) 510-540.

[49] R.W. Vossen, Relative strengths and weaknesses of small firms in innovation, International Small Business Journal, 16(3) (1998) 88-94.

[50] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decision Support Systems, 90 (2016) 23-32.

[51] Z. Xiang, Q. Du, Y. Ma, W. Fan, A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism, Tourism Management, 58 (2017) 51-65.

Chapter 3: Maximizing total yield in safety hazard monitoring of online reviews

Abstract

Many firms face enormous challenges in monitoring their products for evidence of potential safety hazards, which can have profoundly negative effects both on consumers and on firms' financial standings. Recent works have responded to this dilemma by proposing methods utilizing text analytics to rapidly sort and prioritize online reviews. These data sources provide a growing volume of up-to-date feedback, and text analytic methods aim to efficiently sort through this data for vital insights. These methods tend to emphasize precision, or the proportion of retrieved records that reflect true positives. In this paper, we consider cases in which retrieving a greater number of true positives may be an important objective as well, such as when the costs of false negatives are known to be especially severe. To address this problem, we propose several nuanced methods for categorizing these online reviews, including choosing multiple sets of indicative and piecewise "smoke terms" that are predictive of safety hazard-related mentions in online reviews; rank-based smoke term selection; and fuzzy matching. We compare these techniques in multiple product categories and attempt to generalize across categories. We find that these methods can augment existing techniques for detecting mentions of safety hazards in online media, indicating great potential to improve firms' monitoring of consumer feedback and expedite quality-related analytics.

Keywords: text mining, online reviews, safety hazards, business intelligence, fuzzy matching, classification.

1. Introduction

Product safety hazards can have profoundly negative effects on both firms and their consumers. Marucheck et al. [41] describe a myriad of ways in which products have exposed consumers to risk of bodily harm or death, including motor vehicle defects inducing sudden acceleration; tainted food products exposing consumers to food poisoning; and laptop computer batteries that overheat or potentially catch fire. While the specific manner of safety hazard tends to vary by product category, a commonality is that these hazards tend to have dramatic and negative implications for the firms responsible for producing the hazardous products. Firms that are associated with hazardous products often lose the goodwill of their consumers [26], weakening their perception in the market, and studies have also found that these firms suffer from diminished financial performance [51, 60].

While firms are often diligent in their attempts to promote product quality, instances in which unsafe products can make it to market are not uncommon. Many firms, for instance, utilize programs such as Six Sigma to ensure regularity of their manufacturing processes and minimize the rate of defects [61]. Relatedly, firms may perform tests of their products before selling them to consumers to ensure that they perform as intended [49]. However, it is notoriously difficult for firms to exactly replicate the use cases of their consumers, so oftentimes potential flaws are missed in the product testing phase [49]. The United States Consumer Product Safety Commission (CPSC) is tasked with regulating most consumer products in the United States for possible safety hazards, while the Food and Drug Administration (FDA) regulates pharmaceutical products. However, particularly for consumer products, monitoring tends to be reactive rather than proactive. That is, the CPSC responds to reports of clearly hazardous products, but it is unable to preemptively test the myriad of consumer products before they go to

market to ensure their safety. The CPSC runs a National Product Testing and Evaluation Center in Rockville, Maryland, but with so many unique products on the market (for instance, Amazon.com alone sells 564 million unique products [17]), only a small portion can be tested.

Several recent studies (e.g., [1, 3, 22, 43]) have turned to online media as a potential source for intelligence on product safety. In recent years, the volume of online media has expanded dramatically [42]. This phenomenon represents a remarkable opportunity to mine intelligence from the web, as consumers provide a constantly updating picture of their interactions with products. Although the volume of data ensures the availability of vital information, it also complicates the process of extracting this information. As so many online posts exist, it may be incredibly difficult to delineate the most important posts, a phenomenon called information overload [25]. A key area of text analytics focuses on simplifying these problems by classifying text into some number of predefined categories. Applications include classification of spam emails [58], crowdfunding projects [59], and online terror chatter [48]. The literature notes the tension between two competing objectives: precision and recall [19, 52]. Precision refers to the proportion of identified instances that are true positives, while recall refers to the proportion of all positives that have been identified. In general, there is an inverse relationship between these measures; classifiers can improve recall by lowering the threshold of certainty required to label a record as a positive instance, but making less certain designations often sacrifices some precision [45]. F-measure has been proposed as a compromise of these two metrics, which considers a blend of both precision and recall as an overall measure of a classifier's effectiveness [45].

In text analytics, researchers often resolve the conflict between precision and recall in favor of precision. As text data is so voluminous, it is unrealistic in many use cases for

practitioners to read substantial portions of it [1, 3]. Recognizing that only a small portion of the classification may be practically useful, many researchers choose to strive for high precision within that portion [1, 3]. While this approach allows some true positives to escape attention, it ensures an efficient use of resources because any time spent analyzing those records classified as positives is likely useful and productive.

Most recent safety surveillance studies have taken this approach, seeking to maximize the number of true positives observed in a top-ranking portion of online reviews [1, 3, 4, 35, 43, 55]. However, there are several domains in which recall may be a substantial concern as well. Hora et al. [26] describe how different types of safety hazards necessitate unique responses. In particular, the most severe forms of safety hazards, which can cause significant bodily harm or death, necessitate quick and comprehensive responses to protect both social welfare and corporate liability [49]. In other words, the costs of false positives and false negatives are asymmetric; a false negative in this case may be much more problematic than a false positive [31]. While practitioners in these categories might still lack the capability to manually assess all online content, monitoring techniques in these categories may best be designed to consider the depth of solutions, or the total yield of true positives that they return as well as the prioritization of online content (precision). Truly optimizing recall would involve simply reading all online reviews, as this approach ensures that all true positives are identified. As this is likely unrealistic, we consider lift as a more meaningful objective. Lift refers to the ratio of observed prevision to the level expected by random chance. Rather than considering precision only in a top-ranking set of online reviews, we aim to improve lift over the entire distribution of online reviews.

This study examines methods for detection of safety hazard-related discussion in online reviews obtained from multiple product categories. The first category that we consider is the

over-the-counter (OTC) medicine category utilizing online reviews obtained from the world's largest online retailer, Amazon.com. OTC medications, also known as non-prescription medications, are products that consumers can purchase in pharmacies, supermarkets, and online without a doctor's prescription in order to treat common maladies. OTC medication is widespread, and these medications have the capacity to chemically alter body chemistry, so safety hazards have the potential to severely harm or even kill users. Millions of doses of unsafe drugs have appeared on the market before later being recalled [50], implying that prior detection of safety hazards in this category is imperfect. To supplement drug testing, the monitoring of online content would allow drug manufacturers and regulatory agencies to detect unsafe drugs as quickly as possible, mitigating adverse drug reactions and improving the safety of the OTC medicine category. Second, we also apply our technique to online reviews of seasonal items obtained in collaboration with a large Fortune 50 retailer based in the United States. Products classified as seasonal items are only sold for small portions of the year, potentially making it difficult for firms to determine an extensive track record for each product. Additionally, this also represents a diverse product category in that it includes many different types of products. Thus, this product category is quite difficult for retailers to monitor effectively, as they must monitor a wide range of products simultaneously while utilizing information that they have only recently obtained.

In analysis of these product categories, we aim to develop novel techniques for improving safety hazard-monitoring practices. We aim to augment prior approaches in the literature (e.g., [1, 3, 22]) that suggest the use of "smoke terms," or machine-learned terms particularly associated with references to safety hazards within a product category of interest. By utilizing smoke terms, researchers and practitioners can rapidly sort online content such that content

containing more smoke terms is believed to be more likely to refer to safety hazards. Much prior work has been devoted to methods for choosing these terms and comparisons of the performance of smoke terms versus other common text analytics techniques, such as sentiment analysis [1, 3, 22]. In our work, we contribute to this stream of literature by proposing unique approaches that consider this problem from a more lift-focused as opposed to precision-focused perspective. We compare the performance of our approaches to those of prior work and demonstrate their effectiveness and value for our use case. Additionally, as many prior works have limited their scope to a single product category at a time, we also examine the extent to which smoke terms can be applied across product categories and attempt to construct cross-category surveillance mechanisms.

We structure the remainder of this paper as follows. In the next section, we provide a literature review discussing online reviews, contemporary text analytic approaches, statistical classification challenges, and threats of safety hazards in OTC medicine and seasonal items. In the third section, we describe the major research directions that we pursue in this work as well as the contributions of our study. Next, we describe the datasets that we utilize in this study as well as the methods proposed for analyzing these datasets. Utilizing these methods, we detail the results of our techniques in contrast to those used in prior works. We list several limitations in our work. Lastly, we conclude our paper, noting the potential for future work in this research stream and the implications of our study for academia and industry.

2.0 Literature review

In this section, we review prior literature related to our work. We describe past works on online reviews, text analytics, statistical classification challenges, and the prevalence of safety hazards in both OTC medicine and seasonal items. We discuss major findings and methods in contemporary studies, and we also cover key open questions in the literature.

2.1 Online reviews

For firms, a major avenue for interaction with consumers has been in the form of online reviews, through which consumers describe their experiences with a product and detail the manners in which it may have met or failed to meet their expectations [57]. As the Internet has expanded, the volume of online reviews has also substantially grown over time. As of 2013, a single e-commerce retailer, Amazon.com, received over 20 million online reviews within a single year [42], with the rate of growth also rising year to year. Figure 1 displays the growth in volume of online reviews on Amazon.com between 1996 and 2013.

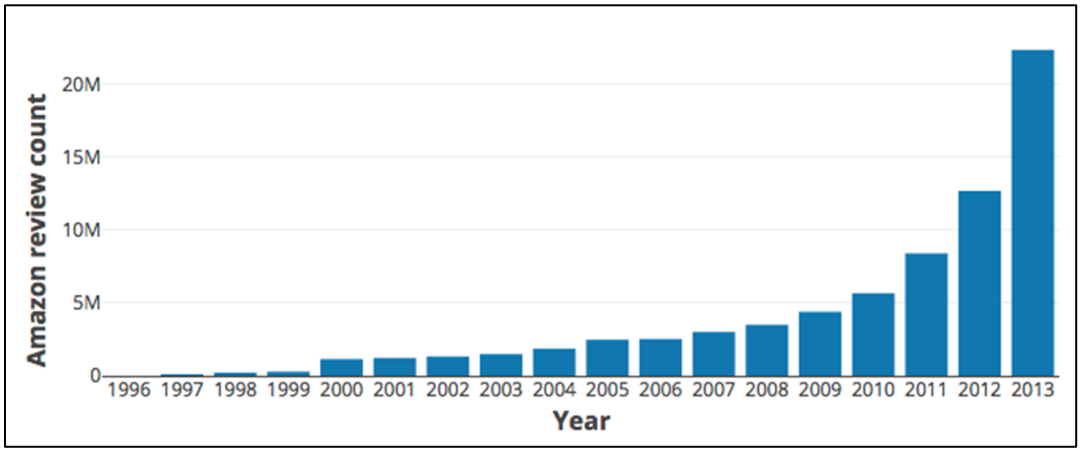

Figure 1. Annual Amazon review count from 1996 through 2013 (derived from data first obtained by McAuley et al. [42]).

The information systems literature has studied online reviews extensively in recent years, as they have profound managerial implications. For example, Chevalier and Mayzlin [9] examined the relationship between star ratings and the subsequent sales of products, finding that products receiving higher star ratings outsold those with lower star ratings. While more positive reviews are intuitively thought to be more desirable for firms to receive than more negative reviews, some research has argued that simply receiving reviews is valuable to a firm because the reviews offer its products credibility [7]. Surveys have found that consumer usage of online reviews is widespread, as 91% of consumers report reading online reviews before making purchasing decisions [8]. Importantly, the literature has suggested value in firms being responsive to their online reviews [46], as firms that are aware of their consumers' expectations can more easily adapt to meet them. The literature has utilized online reviews as a data source for many different areas, including forecasting demand [6, 33], deriving market intelligence on competing products [38, 57], and safety surveillance [1, 3, 22, 43].

2.2 Text analytics

The Internet has been a vibrant medium for interpersonal communication, allowing individuals to express their opinions for anyone else in the world to view. Text analytics has become a popular area of research in recent years as the spread of Internet connectivity has led to a massive quantity of text becoming available online. Online reviews are a key source of text for firms to utilize, but other valuable sources of text include social media, blogs, and news articles [13, 15]. Text data can be quite interesting as it allows a user to express a rich and nuanced message; however, these data sources are also difficult to analyze because they are unstructured (that is, it the data does not conform to easily identifiable fields) [13, 15]. In some cases,

algorithms may be used to manipulate and/or simplify text so that it is more manageable for ensuing analytical techniques.

Previous works have explored many different angles of analyzing textual data. Sentiment analysis or opinion mining refers to techniques that attempt to quantify the emotive content expressed in text [27, 44]. Named entity recognition refers to identification and extraction of entities described in text, such as people, places, or organizations [47]. Topic modeling is used to cluster text into dominant topics or themes of discussion [11]. Text classification refers to efforts to separate textual documents into different categories [19, 52]. These categories may be predefined in supervised classification, or a computer algorithm may define the categories itself in unsupervised methods.

Text classification has been of interest for recent efforts in safety surveillance, as text can be classified based on whether it mentions a safety hazard that would be important for a firm or regulator to review. As these categories are pre-defined, supervised classification has generally been preferred [1, 2, 22]. Recent literature has found that an effective mechanism for text classification is to use smoke terms, or terms especially associated with mentions of safety hazards, to rank online content from most likely to least likely to refer to safety hazards [1, 3, 22]. The top-ranking portion of content is thought to refer to safety hazards, and it can be subjected to further analysis by stakeholders. Unfortunately, these terms tend to be category-specific, as terms like “airbag” may be highly indicative of safety hazard mentions in vehicles, but they have almost no use in pharmaceuticals [3]. Prior work has identified smoke terms suited for a variety of product categories, including vehicles [3], dishwashers [35], toys [55], baby cribs [43], and countertop appliances [22].

2.3 Statistical classification challenges

In classification, precision (or sometimes confidence) refers to the proportion of retrieved records that are true positives, and recall (or sometimes sensitivity) refers to the proportion of all positive records that have been retrieved [19, 52]. Ideally, a classifier would have high recall and high precision, identifying a set of records as predicted positives that are largely true positives and that represent a large portion of all positives. Often, however, the user experiences a trade-off between the two measures, as techniques that are particularly sensitive will have higher recall because they identify more records as predicted positives but lower precision because a smaller proportion of those records are true positives [19, 52]. Conversely, a technique may have a high threshold for a predicted positive, in which case precision would be higher because most predicted positives would be true positives, but recall would be lower because a smaller proportion of all positives would be identified [19, 52]. In this work, we also consider lift, or the ratio of observed precision to the level of precision that would be expected by random chance. Incorporating lift, the user can determine how many records to review, and given that number of records, lift reports how effectively the user's time is spent.

These competing measures are of vital importance in safety surveillance. Often, an argument is made in favor of precision: as only a small portion of online content refers to safety hazards, an efficient search is necessary to derive meaningful intelligence [22]. Yet, the costs of misclassification are asymmetric [31], as a false positive is relatively inexpensive, but a false negative can lead to harm for consumers and financial responsibility on the part of the firm. In past work, Goldberg and Abrahams [22] proposed a heuristic method for improving precision in a set of top-ranking reviews, which has been extremely effective. However, as the high-precision smoke term lists tend to be rather short, they are most effective in this top-ranking range. A

limitation of this study is that the smoke term lists curated by heuristic methods do not offer a substantial depth of solution; they sacrifice recall in favor of precision. Figure 2 displays a lift chart to demonstrate this point; the smoke term list is highly effective for the top few hundred documents, but after exhausting documents that contain those terms, the remainder of the documents are classified at the level of random chance. Techniques that incorporate a larger portion of records are still elusive in the safety surveillance literature.

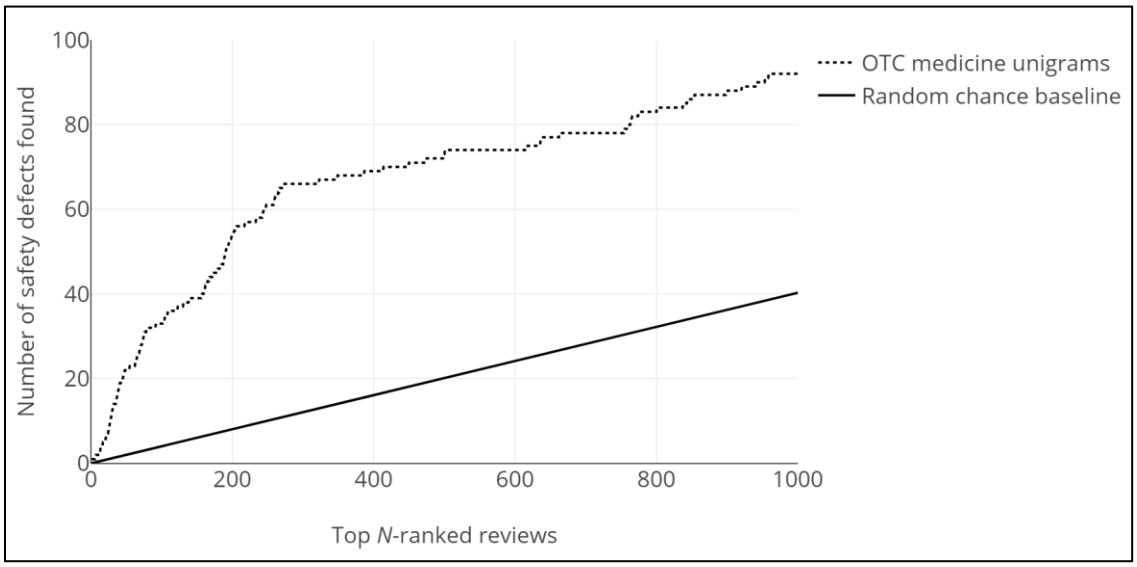

Figure 2. Lift chart of OTC medicine smoke term unigrams (adapted from Goldberg and Abrahams [22]).

2.4 Safety hazard concerns

2.4.1 OTC medicine and pharmacovigilance

OTC medication safety is a concern to many stakeholders, including consumers, drug manufacturers, wholesale drug distributors, re-packagers, dispensers (primarily pharmacies), and government regulators. In the past decade, many medications that cause severe adverse

symptoms (e.g., nausea, vomiting, seizures, loss of consciousness, dizziness, and fatigue) have been recalled by the FDA [50]. Pharmacovigilance refers to efforts to detect and prevent adverse drug reactions. Pharmacovigilance systems have been instituted to both by regulators, such as the FDA's Adverse Event Reporting Systems (FAERS), and by industry advocates, such as MedDRA [4]. However, while pharmaceutical firms are required to report adverse reactions during trials, reporting after drugs have been released is generally voluntary [12, 23], and adverse reactions may be seriously underreported [24, 40]. The medicinal industry is also difficult to regulate due to a tension between external pressures and safety concerns: industry stakeholders and consumers often push for medications to be released to market as quickly as possible, but doing so can result in unsafe products making it to market [56]. In this category, safety hazard-related discussions are extremely valuable for stakeholders to read, as adverse reactions can permanently harm consumers. As such, the goal of detecting as many hazard-related discussions as possible is paramount.

2.4.2 Seasonal items

Seasonal items are a notoriously difficult product category for retailers to manage, as the products are quite diverse and are only sold for short periods of the year [53]. A consequence is that retailers may not have a long or reliable track record to depend upon for vetting each potential product, exposing the retailer to greater risk that a product is of poor quality and/or hazardous [53]. As a result, safety hazards may be overlooked in seasonal items; for example, Weidenhamer [54] reported instances of lead contamination across multiple series of holiday-related products. Further complicating matters, retailers are often concerned with the uncertainty in forecasting demand for seasonal items, so they often purchase the products from wholesalers

as close to the time of sale as possible [30], minimizing the time available to thoroughly vet the quality of these products. Any ability to more rapidly source intelligence on these seasonal items would improve retailers' capacities to manage this difficult product category more effectively. Given the diversity of the products in this category and the relatively short track record for each product, safety surveillance is especially challenging, and detection of as many hazard-related discussions as possible is necessary to ensure that dangerous products are quickly removed from the market.

3.0 Research direction and contribution

This paper aims to demonstrate several possible methodologies for improving the depth of the solutions offered by term-based classifiers, increasing the total yield of true positives detected. In this paper, these techniques will be applied to safety hazard detection in the OTC medicine product category and the seasonal items product category; however, these techniques are also applicable for additional use cases for which improving the yield of true positives is also desirable. Typically, for these applications, true positives are very useful, and false negatives are very harmful, such as the detection of terror chatter from online discussions [48].

We seek to address this problem from multiple angles. First, rather than solely considering precision in top-ranking reviews as the metric for a high-performing method, we propose two alternative objective formulations, one based on a piecewise function and one based on the sum of ranks, for term selection. Second, we suggest the use of fuzzy term matching to improve recall by accounting for misspellings, inflected forms of terms, or other variations. We show the application in two large datasets from different product categories. Third, we also look for commonalities between smoke terms generated for our product categories of study and those studied in prior works, testing the generalizability of these techniques.

The first major contribution of this paper is to present new text analytic techniques for the purposes of more depth-driven surveillance. Such text analytic techniques are not as widely studied as purely precision-focused techniques [19, 52], so improving these methodologies addresses a key gap in the literature. We measure this performance mainly in terms of lift, as the user can specify an arbitrary cutoff at which we compare performance to random chance. Thus, we consider occasions in which a user may wish to examine a greater proportion or reviews. Although we apply our techniques for safety surveillance in this paper, these techniques would

also be applicable in other domains, such as detection of terror chatter [48]. Furthermore, this paper suggests several heuristic methods for addressing text analytics problems; uses of management science techniques in text analytics are still rather uncommon, and hopefully these techniques will spur on further exploration of complementary analyses.

Second, for practitioners, this paper will offer usable text analytic insights for firms and regulators in the OTC medicine and seasonal items product categories. For stakeholders interested in these product categories, these new tools and insights provide improvements to the safety surveillance process with which to source actionable safety-related insights.

Third, comparison of our results to prior works and construction of cross-category tools has substantial implications for safety surveillance. The construction of these supervised machine learning methods can require a great deal of setup as records must be manually labeled to train these techniques. While we expect category-specific techniques to offer the best performance, a high-performing cross-category tool would improve surveillance for product categories for which these more specialized techniques do not currently exist.

4.0 Methodology

4.1 Dataset and data coding

In this study, we make use of two key datasets, which demonstrate the effectiveness of our techniques in multiple contexts. First, the research will make use of online review data from Amazon.com, the world's largest e-commerce retailer and the most voluminous online review platform [42]. 12,400 OTC medicine reviews posted on Amazon.com between 2008 and 2013 were randomly selected for analysis in this study, including allergy medication, cough syrup, antacids, digestion aids, and pain relief medication. Second, in collaboration with a large Fortune 50 retailer based in the United States, we obtained a further 36,488 reviews of seasonal items posted on that firm's online review platform in 2017 and early 2018. As these reviews spanned just over one year, they pertain to a variety of mixed products varying in accordance with weather, holidays, and promotions. Further details on each dataset are shown below in Table 1.

Table 1. Descriptive statistics on online review datasets.

	OTC medicine (Amazon.com)	**Seasonal items (Fortune 50 retailer)**
Count of reviews	12,400	36,488
Count of unique products	1,028	5,271
Count of unique manufacturers	203	647
Date range	01/01/2008 to 12/09/2013	01/01/2017 to 03/20/2018
Review-level character count (min / mean / median / max)	5 / 345.9 / 221 / 15,096	2 / 387.4 / 268 / 8,844
Review-level word count (min / mean / median / max)	1 / 63.8 / 41 / 2,755	1 / 73.2 / 51 / 1,568
Count of unique *n*-grams (unigrams / bigrams / trigrams)	16,243 / 189,289 / 466,251	26,413 / 507,135 / 1,488,326
Mean / median star rating	4.4 / 5	3.6 / 4
Percentage 1-star reviews	7.5%	23.7%
Percentage 2-star reviews	4.0%	6.8%
Percentage 3-star reviews	6.1%	7.2%
Percentage 4-star reviews	14.8%	15.2%
Percentage 5-star reviews	67.6%	47.1%

Next, we devised a protocol for dividing these online reviews into two mutually exclusive classes, “safety hazard” and “no safety hazard,” based on the examples from prior work [1, 3, 22] and adapted most closely from Goldberg and Abrahams [22]. We delineate between these two classifications as follows:

1) “Safety hazards” refer to reviews in which the consumer indicates a serious problem with a product that either has already caused or has the potential to cause some sort of bodily harm, death, and/or property damage. In OTC medicine, safety hazards typically refer to adverse reactions to drugs, with effects potentially including vomiting or seizures. Safety hazards can manifest in a variety of ways depending on the seasonal items; some examples include products shattering and exposing the consumer to broken glass or products catching fire and exposing the consumer and their property to risk of burns. The following is an example of a seasonal item determined to reflect a safety hazard:

> *“Dangerous! DO NOT BUY THIS CHAIR. We have had 3 of these chairs completely collapse without warning on average-sized adults. The last time it happened, the plastic from the chair splintered, causing a painful laceration to my husband's hip area in which he still has a scar.”*

2) “No safety hazards” refer to any reviews that do not meet the qualification for “safety hazard” discussed above. Reviews that give general information on the product’s performance or quality not related to safety fall under this category whether the content is positive or negative. For example, a consumer could complain that their medication was not an effective decongestant, but this would not suggest a safety hazard. Alternatively, some reviews just gave general comments or emotion that did not specify aspects of a product’s performance, and such

reviews are also considered not to be safety hazards. The following is an example of a seasonal item determined to reflect no safety hazard:

> *"looks nice in my yard. I bought the doe and the raindeer and loved them both. The bow is big and I thought it was going to be hard to put together, but it was a breeze. And with the stakes to go in the ground, it stayed put even through the wind."*

Following the above protocol, we tasked groups of undergraduate business students at a large public research University based in the United States with labeling or "tagging" reviews as either "safety hazards" or "no safety hazards." To alleviate bias, reviews were assigned to students at random. We asked each tagger to tag no more than 200 reviews to ensure that quality was not compromised due to fatigue. 107 students were assigned to tag OTC medicine reviews, and 197 students were assigned to tag seasonal item reviews.

The student taggers completed 13,794 tags spanning 10,874 OTC medicine reviews. Due to random assignment of taggers to reviews, some reviews were tagged multiple times. When multiple students tagged the same review, they were unanimous 2,110 times (97.8% of cases) and disagreed just 47 times (2.2% of cases). To further verify the quality of these tags, the lead author completed 300 tags as a means of "authority tagger" comparison. Student taggers agreed with the authority tagger 91.3% of the time. Cohen's κ [10] compares this value to random chance, and in this case we observed a value of 0.83. This level of agreement is acknowledged by Landis and Koch [34] as "almost perfect" and by Fleiss et al. [18] as "excellent."

For seasonal items, the student taggers completed a total of 42,990 tags spanning the 36,488 reviews. When multiple students tagged the same review, they were unanimous 3,851 times (91.8% of cases) and disagreed just 346 times (8.2% of cases). On this project, two graduate students were assigned as authority taggers for comparison to the undergraduate

taggers, and the graduate students tagged a total of 1,079 reviews. The rate of agreement between the student taggers and authority taggers was 86.3%, yielding a Cohen's κ [10] value of 0.73. Landis and Koch [34] rate this level of agreement as "substantial," and Fleiss et al. [18] rate this level of agreement as "fair to good."

For both product categories, the levels of agreement both among student taggers and between student taggers and authority taggers provide convincing evidence that the tagging protocol was applied consistently and properly, and thus the final decisions are of high quality. In the rare cases of disagreement between our taggers, we resolved the disagreements using a "most conservative" decision rule, always classifying these reviews as safety hazards. Winkler et al. [55] and Goldberg and Abrahams [22] each argue that the cost of a false negative is far higher in safety surveillance than a false positive; thus, we assume that these reviews refer to genuine safety hazards to avoid missing potentially important safety hazard-related reviews in our analysis. In both product categories, the final rate of safety hazards observed was approximately 2 percent.

4.2 Data processing

Having tagged both datasets of reviews, we next employ a variety of data processing steps to build mechanisms for automated ranking and sorting of reviews. Prior work has extensively examined the use of smoke terms, or words and phrases particularly prevalent in and predictive of safety hazard-related reviews, for similar problems [1, 3, 22, 39]. A major challenge in this technique, however, is the choice of smoke terms. Due to the complexity and diversity of language, even small numbers of records tend to contain numerous possible unigrams (single words), bigrams (two-word phrases), and trigrams (three-word phrases). In past

works, the numbers of candidate smoke terms have been in the thousands or even millions [22], making the selection of appropriate smoke terms a difficult problem. Per Table 1, the same is true of our datasets.

As an initial filtering mechanism, prior work (e.g., [1, 22, 35, 55]) suggests the use of Fan et al.'s [14] CC score, an information retrieval technique which uses the X^2 distribution to provide relevance scores for candidate terms (*n*-grams). Higher scores imply that the term occurs quite frequently in the positive class and quite infrequently in the negative class. As such, terms with high scores may be good predictors of the positive class. Although useful for identifying candidate smoke terms, the technique is insufficient on its own, as overfitting to the training set may be problematic, and the technique also does not account for the interplay between different terms together.

We define the term selection problem as follows, and a brief schematic of this procedure is displayed in Figure 3. We divide our initial dataset into three segments of equal size: a training set (A), a curation set (B), and a holdout set (C). The training set is used to gather initial candidate terms using Fan et. al [14]'s CC score (1). The curation set is used to test the efficacy of various combinations of smoke terms and derive a high-performing list (2). Finally, the holdout set is used to evaluate and benchmark these smoke terms on unseen data (3). For benchmarks, we compare the efficacy of our smoke terms to common sentiment analysis approaches, namely AFINN [44] and Harvard General Inquirer [32], and to a random chance baseline.

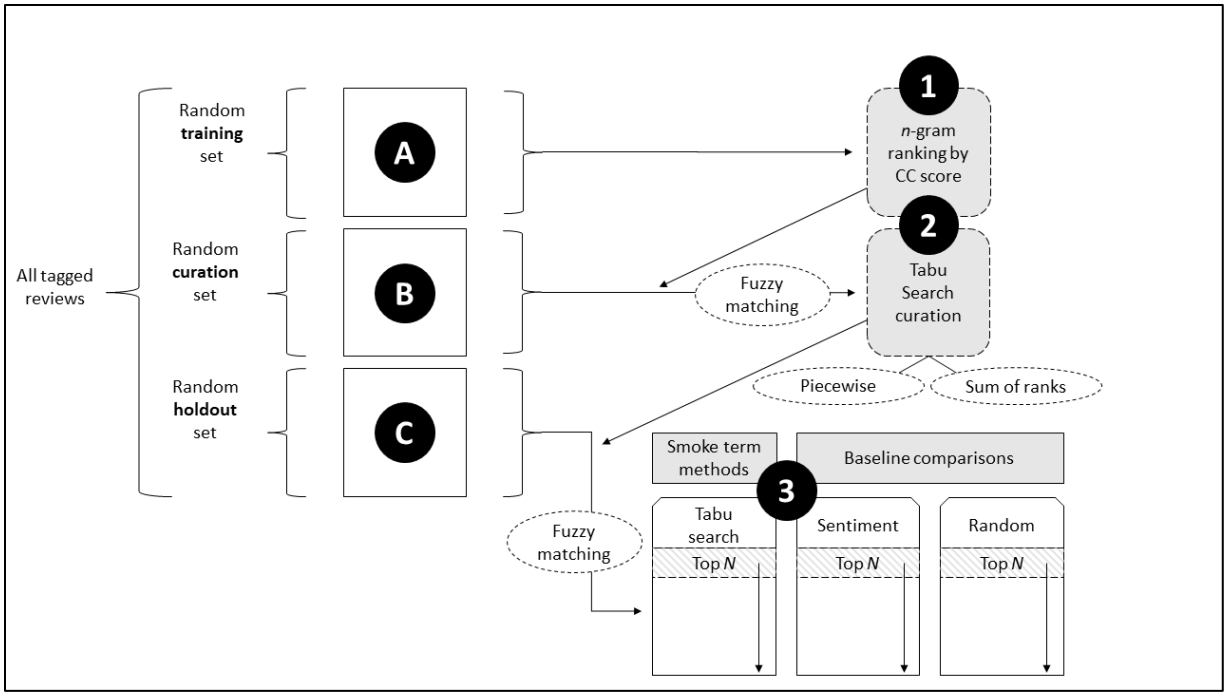

Figure 3. Schematic of text analytic methodology.

We denote each term as t in the initial set of the candidate smoke terms. We create a binary variable for each term, x_t, which equals 1 if term *t* is included in the term list and 0 otherwise. We represent the chosen solution with a vector, S, which consists of the set of variables x_t for $t = 1 \dots T$, or $[x_1\ x_2\ x_3 \dots\ x_T]$. Each smoke term list can be used to provide a score for a given review, where a higher score indicates a greater likelihood that the review refers to a safety hazard. These scores are calculated using the count of the occurrences of each term (from a document-term matrix) weighted by the terms' CC scores [3, 22]. Goldberg and Abrahams [22] address the term selection problem by seeking to maximize a weighted precision score for several cutoffs of the top *N*-ranked reviews. As smoke terms are intended to assist with prioritizing a series of documents through ranking, we use the function $f(S, N)$ to denote the number of true positive documents found in the top N-ranked reviews of the curation set based on the smoke term list S. The function $f(S, N)$ presents a difficult optimization problem because

it involves *ranking* a set of reviews using the smoke term list S. Consider some candidate terms for the OTC medicine product category. For example, the term list ["vomit"] may yield 20 true positives out of the top 100-ranked reviews, and the term list ["dangerous"] may yield 15 true positives out of the top 100-ranked reviews. However, combining these smoke term lists to form ["vomit", "dangerous"] may only yield 18 true positives out of the top 100-ranked reviews.

Prior research has suggested the use of a Tabu search heuristic for term selection [21, 22]. For this heuristic, each candidate smoke term list is considered a possible solution. This heuristic seeks to maximize an objective function (in this case, the number of true positives found in the top N-ranked reviews) by examining and testing solutions neighboring the current solution [20]. The underlying component of the heuristic is greedy, always pivoting to the neighboring solution that provides the greatest improvement in objective function. Unlike a pure greedy heuristic, if no improving move is found, the Tabu search allows movements in directions that decrease the objective function so that more of the solution space is explored. Previously explored solutions are remembered to ensure that the algorithm does not cycle perpetually.

Although this technique has been largely effective in producing lists of smoke terms that provide excellent performance for precision, these lists may not provide desirable levels of recall. Thus, in the following, we suggest several methodological advances or alternatives for use cases in which recall is also an important consideration.

4.2.1 Piecewise Tabu search-curated smoke terms

In using a precision-based Tabu search technique [21, 22], the user sets a cutoff for the top N-ranked reviews, N. We define our nonlinear optimization problem as follows.

$$Maximize\ f(S, N)$$

$$x_t\ binary\ \forall\ t$$

This initial technique maximizes precision in the top-ranking set of reviews using a Tabu search heuristic. To create piecewise smoke term lists, after generating this initial set of smoke terms, we remove the top N-ranked reviews as specified by the user's cutoff from the curation set. Then, the Tabu search algorithm is run again to generate a new set of terms that maximize precision in the next set of top-ranked reviews. In effect, this algorithm trains several smoke terms lists that each maximize precision. If a review receives a positive score via the first smoke term list, then its score is retained; however, if not, then the second smoke term list is run to see if it scores positively via that list. Each review is run through multiple lists to check for hazard-identifying terms, which in effect increases the number of true positives found. In sum, recall may be improved by expanding the number of precision-based lists in use. Over the whole of the corpus, we expect superior lift as the additional smoke term list(s) will match a greater number of positives.

4.2.2 Sum of ranks Tabu search objective

In improving the depth of a solution, another approach is to attempt to systematically shift true positive reviews as much as possible toward the top-ranking reviews. While the former technique emphasized the top *N*-ranked reviews, it does not consider any reviews ranked $N + 1$ or higher. In this formulation, we seek to minimize the ranks of the true positive reviews given the chosen ranking system, as lower ranks imply that these reviews are sorted to the top of the list. Rather than focusing only on a top-ranking set of reviews, the entire distribution of reviews

is considered in this approach. We denote each document (review) in the corpus as d_i, indexed from $i \ldots k$, and the rank of each review given the smoke term list S as $r(d_i, S)$. Therefore, in this technique, we seek to minimize the following objective.

$$Minimize \sum_{i}^{k} r(d_i, S)$$

$$x_t \; binary \; \forall \; t$$

The minimization of this objective weights all reviews equally and attempts to shift each true positive review as far as possible towards the front of the distribution (lower ranks). The Tabu search heuristic is still applicable to this objective, as it is still attempting to optimize a ranking function, although lower values are preferable for this function. This formulation is consistent with improving lift, as the entire corpus of reviews is considered jointly. Lowering the value of the objective implies that true positives are shifted towards higher ranks, improving lift at any arbitrary cutoff.

A possible extension of this approach is to apply some weighting system to the rankings such that poor rankings are especially penalized. We define w as a user-specified weight, where $w \geq 1$. We apply w as an exponent to penalize worse ranks in our objective function as follows:

$$Minimize \sum_{i}^{k} r(d_i, S)^w$$

$$x_t \; binary \; \forall \; t$$

The choice of $w = 1$ is equivalent to the previous formulation in that the ranks are weighted linearly in the objective function. However, if $w = 2$, for example, then worse ranks

would be penalized via a quadratic function. We do not pursue this extension in the scope of our paper, or in essence we use $w = 1$.

4.2.3 Fuzzy term matching

To the best of our knowledge, prior research on smoke terms has used exact term matching to score documents [1, 3, 22]. That is, if a document contains the exact text of the smoke term, then it is scored positively, but any spelling variations or non-identical phrasing are not caught. In this work, to improve recall, fuzzy term matching will be used to increase the number of reviews that contain each smoke term. There are several methods for fuzzy term matching, although Levenshtein distance [36] is among the most popular. Levenshtein distance assesses string similarity according to the following function, which measures the distance between the first i characters of string a and the first j characters of string b, where $1_{a_i \neq b_j}$ is 1 when $a_i \neq b_j$ and 0 otherwise.

$$LD_{a,b}(i.j) = \begin{cases} \max(i,j) \\ min \begin{cases} LD_{a,b}(i-1,j) + 1 \\ LD_{a,b}(i,j-1) + 1 \\ LD_{a,b}(i-1,j) - 1 + 1_{a_i \neq b_j} \end{cases} \end{cases}$$

The output of this function may best be understood as an "edit distance," or in other words the minimum number of edits required to transform one string into another, whether the possible edit operations are substitutions, deletions, or insertions of one character. For example, the distance between "kitten" and "mitten" is 1 because the substitution of a single character is required to transform one string into the other. Using a threshold of 1, for example, would require near-exact matches, whereas a threshold of 3 would not require matches to be as close,

potentially introducing more false positives. In this study, we will show results from several possible thresholds. In the case of phrases (bigrams or trigrams), we apply the Levenshtein distance to the entire phrase. For example, the distance between “after 4 hours” and “after 5 hours” is 1.

In this paper, we examine the impact of fuzzy matching in two distinct senses. The first sense examined in our experiments is the use of fuzzy matching in smoke term curation. In addition to the aforementioned techniques, we also attempt the same techniques while incorporating fuzzy matching as opposed to exact matching. We compare the terms selected with or without fuzzy matching as well as the performance of those terms. The second sense examined in our experiments is the use of fuzzy matching purely for identifying safety hazards within our holdout set. In these experiments, we do not utilize fuzzy matching for term selection, but we do apply it for the ranking and prioritization of our dataset. We discuss findings from this array of experiments in the following section.

5.0 Results and evaluation

5.1 Heuristic term selection and performance

We display the terms selected by our techniques for each product category in Tables 6-11 of Appendix A. For piecewise Tabu search-curated smoke terms, first-tier refers to the set of terms curated by the first Tabu search, and second-tier refers to the set of terms curated after removing the top *N*-ranked reviews from the curation set and rerunning the algorithm. We found that the first two tiers of smoke terms were useful, but after this point, too few hazards remained in the curation set to generate further meaningful smoke term lists. We also display terms curated using the sum of ranks Tabu search, and finally we indicate whether each term was identified in any prior work(s) utilizing smoke term methodologies.

Interestingly, we observed substantial concordance between the piecewise smoke term technique and the sum of ranks Tabu search technique. Typically, the sum of ranks Tabu search technique spanned most of the terms contained together between the first-tier and second-tier piecewise Tabu search smoke term lists. However, in each case, the sum of ranks Tabu search technique omitted several terms that the piecewise Tabu search smoke term list obtained while including several terms not found in the piecewise Tabu search smoke term list.

In Figure 4, we display lift charts that visually show the effectiveness of each technique in ranking the reviews in the holdout set from most likely to least likely to refer to a safety hazard. For any arbitrary cutoff of the top *N*-ranked reviews on the x-axis, the y-axis shows the number of safety hazards detected. In each chart, the first-tier Tabu search and piecewise Tabu search yield identical curves until the first-tier Tabu search exhausts matches, at which point the piecewise Tabu search overtakes it. We denote this point in each lift chart using a vertical dotted line.

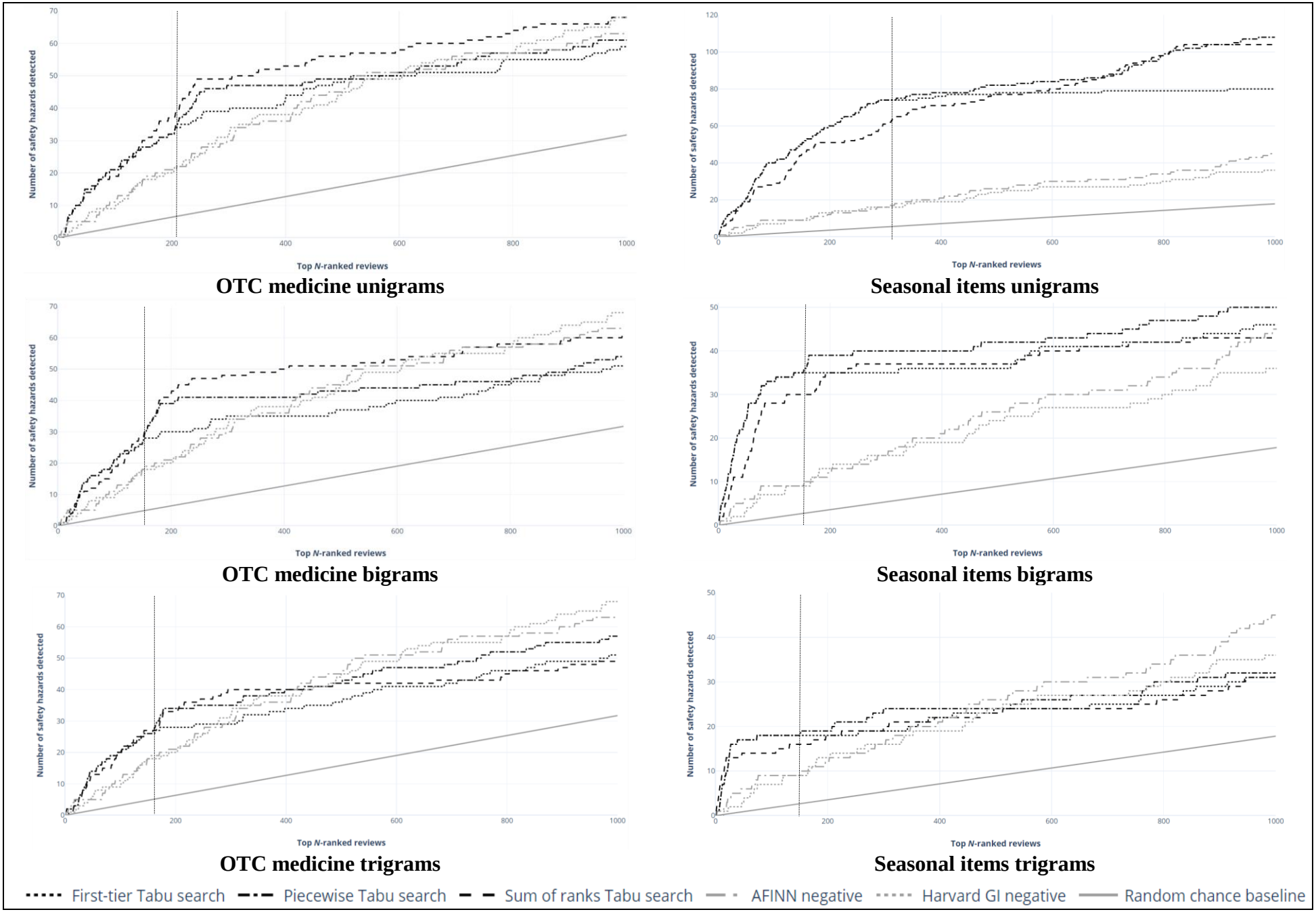

Figure 4. Lift charts comparing performance of each technique.

In general, it appeared that each of the smoke term methods outperformed the sentiment baseline methods (AFINN [44] and Harvard General Inquirer [32]), and all text analytic methods outperformed random chance. Our unigram methods tended to be the most effective of the smoke term methods at detecting safety hazards. It appeared that the piecewise Tabu search was a viable alternative relative to running a single tier of the Tabu search algorithm for smoke term curation. A single tier produces a high-performing set of terms that are highly predictive of safety hazards, reflected by a substantial "bump" in the lift chart. However, after all reviews with matching terms are exhausted, the remainder of safety hazards are detected at the rate of random chance. For the piecewise Tabu search, the lift charts essentially display a small second bump reflecting the performance of the additional smoke term list. Thus, this technique improves the depth of solution by increasing the number of reviews that can be analyzed at better than random chance, and it does so with no reduction in performance over the initial smoke term list.

The sum of ranks Tabu search typically had similar performance to the piecewise Tabu search. The series for the sum of ranks Tabu search typically did not have as pronounced bumps as those of the piecewise Tabu search; however, the smoother curve often behaved very similarly otherwise. This performance is consistent with the terms curated in the sum of ranks Tabu search lists (see Appendix A). These terms were largely consistent with those generated over both piecewise Tabu search tiers. Interestingly, while the sum of ranks Tabu search slightly underperformed versus the piecewise Tabu search in seasonal items, it slightly outperformed the piecewise Tabu search in OTC medicine for some cutoffs. The smoke term methods offered improvement relative to sentiment analysis or random chance, but the choice of the highest-performing smoke term method depends upon the specific application. Particularly for the trigram lists, the depth of solution suffered due to the specificity of terms, and sentiment was a

superior option if the user chose to analyze a very large number of reviews (nearly 1,000). We note that each of these commonly-used sentiment methods assesses sentiment at the word (or unigram) level. As these methods assess a wide variety of words, it is expected that sentiment may offer reasonably a great deal of recall but potentially poor precision.

In Table 2, we report scores for precision, recall, and lift for each of our techniques at cutoffs of the top 100, top 200, top 500, and top 1,000 reviews. The highest-performing techniques are **bolded**. The findings from Table 2 largely echo those from the lift charts in Figure 4. We observed that the unigram smoke term methods were the highest-performing. The piecewise Tabu search was a compelling option at each cutoff as it tied with the first-tier Tabu search for the best precision at lower cutoffs; yet, at higher cutoffs, it offered improved solutions due to superior depth. Particularly within the OTC medicine product category, however, the sum of ranks Tabu search offered competitive performance and in some cases higher performance if a sufficiently large cutoff was chosen.

Table 2. Precision/recall/lift over the top *N*-ranked reviews for each technique.

		OTC medicine				Seasonal items			
Cutoff		**Top 100**	**Top 200**	**Top 500**	**Top 1,000**	**Top 100**	**Top 200**	**Top 500**	**Top 1,000**
Unigrams	**First-tier Tabu search**	**0.300** **0.130** **9.457**	**0.210** **0.183** **6.620**	0.160 0.278 5.043	0.096 0.417 3.026	**0.400** **0.183** **22.444**	**0.300** **0.275** **16.833**	0.156 0.358 8.753	0.080 0.367 4.489
	Piecewise Tabu search	**0.300** **0.130** **9.457**	**0.210** **0.183** **6.620**	0.160 0.278 5.043	0.098 0.426 3.089	**0.400** **0.183** **22.444**	**0.300** **0.275** **16.833**	**0.164** **0.376** **9.202**	**0.108** **0.495** **6.060**
	Sum of ranks Tabu search	0.280 0.064 8.826	0.190 0.087 5.989	**0.185** **0.170** **5.832**	**0.112** **0.257** **3.530**	0.280 0.128 15.711	0.255 0.234 14.308	0.154 0.353 8.641	0.104 0.477 5.835
Bigrams	**First-tier Tabu search**	0.280 0.122 8.826	0.210 0.183 6.620	0.150 0.261 4.728	0.074 0.322 2.333	0.330 0.151 18.516	0.175 0.161 9.819	0.072 0.165 4.040	0.046 0.211 2.581
	Piecewise Tabu search	0.280 0.122 8.826	0.210 0.183 6.62	0.195 0.339 6.147	0.086 0.374 2.711	0.330 0.151 18.516	0.195 0.179 10.941	0.084 0.193 4.713	0.050 0.229 2.806
	Sum of ranks Tabu search	0.220 0.096 6.935	0.210 0.183 6.620	0.150 0.261 4.728	0.074 0.322 2.333	0.280 0.128 15.711	0.175 0.161 9.819	0.074 0.170 4.152	0.043 0.197 2.413
Trigrams	**First-tier Tabu search**	0.280 0.122 8.826	0.210 0.183 6.620	0.195 0.339 6.147	0.086 0.374 2.711	0.100 0.083 10.100	0.090 0.083 5.050	0.048 0.110 2.693	0.031 0.142 1.739
	Piecewise Tabu search	0.280 0.122 8.826	0.200 0.174 6.304	0.170 0.296 5.359	0.084 0.365 2.648	0.170 0.078 9.539	0.095 0.087 5.330	0.048 0.110 2.693	0.032 0.147 1.796
	Sum of ranks Tabu search	0.260 0.113 8.196	0.200 0.174 6.304	0.165 0.287 5.201	0.084 0.365 2.648	0.140 0.064 7.855	0.090 0.083 5.050	0.046 0.106 2.581	0.032 0.147 1.796
Baseline	**AFINN negative**	0.100 0.040 3.152	0.110 0.096 3.467	0.105 0.183 3.310	0.090 0.391 2.837	0.090 0.041 5.050	0.060 0.055 3.367	0.052 0.119 2.918	0.045 0.206 2.525
	Harvard GI negative	0.120 0.052 3.783	0.090 0.078 2.837	0.100 0.174 3.152	0.090 0.391 2.837	0.070 0.032 3.928	0.065 0.060 3.647	0.048 0.110 2.693	0.036 0.165 2.020
	Random chance baseline	0.032 0.014 1.000	0.032 0.028 1.000	0.030 0.055 1.000	0.032 0.138 1.000	0.018 0.008 1.000	0.018 0.016 1.000	0.018 0.041 1.000	0.018 0.082 1.000

In Table 3, we also report Area Under the Curve (AUC) statistics [16, 29] for each method, **bolding** the top-performing techniques. AUC assesses the entirety of the lift curve for each technique by scaling the geometric area under that curve from 0 to 1. The best models produce many true positives in top-scoring items and achieve values close to 1; the worst models produce many false positives in top-scoring items and achieve values close to 0; and 0.5 is the random chance baseline. The AUC statistics largely confirm the results from the previous table. The unigram approaches typically outperformed the alternatives, and the piecewise Tabu search always outperformed the first-tier Tabu search. For OTC medicine, the sum of ranks Tabu search outperformed the piecewise Tabu search for unigrams and bigrams, although the piecewise Tabu search performed better in all other cases. We observed that the sentiment approaches were competitive with the trigram techniques, but the unigram and bigram techniques each outperformed sentiment.

Table 3. AUC statistics for each technique.

		AUC	
Method		**OTC medicine**	**Seasonal items**
Unigrams	**First-tier Tabu search**	0.635	0.675
	Piecewise Tabu search	0.681	**0.696**
	Sum of ranks Tabu search	**0.735**	0.694
Bigrams	**First-tier Tabu search**	0.632	0.593
	Piecewise Tabu search	0.639	0.625
	Sum of ranks Tabu search	0.694	0.569
Trigrams	**First-tier Tabu search**	0.614	0.548
	Piecewise Tabu search	0.661	0.578
	Sum of ranks Tabu search	0.619	0.548
Baseline	**AFINN negative**	0.624	0.619
	Harvard GI negative	0.637	0.614
	Random chance baseline	0.500	0.500

5.2 Fuzzy matching and performance

In addition to the analyses of various heuristic approaches to smoke term selection, we also experimented with fuzzy matching utilizing Levenshtein distances. The prior analyses all used exact matching (i.e., Levenshtein distance of 0) in both the curation step and in the evaluation of the holdout set. We next discuss the potential value of incorporating this fuzzy matching technique in either part of the analysis.

Repeating the curation steps for each technique utilizing various cutoffs for Levenshtein distances (edit distances within 1 or 2), we found that the smoke term lists curated by each heuristic method when analyzing within a Levenshtein distance of 1 were nearly identical to those curated when utilizing exact matching. A Levenshtein distance of 1 implies a very close match between two terms (they only differ by a single character), so the number of matching records only increases slightly when incorporating this method, hence the very similar performance. In Table 4, we provide an example contrasting the term selection in the seasonal items product category with and without incorporation of Levenshtein distances (abbreviated LD). In this example, only a single term differs ("finger" is excluded) once fuzzy matching is incorporated.

When curating terms within a Levenshtein distance of 1, we found that lists most often differed from exact matching by a single term; in some cases, they did not differ at all, and at most they differed by just two terms. When increasing the threshold for Levenshtein distance to 2 or beyond, we found that performance suffered dramatically. Ultimately, many more potential terms were within these edit distances than within an edit distance of 1, and these additional matches tended to introduce more signal than noise. With these larger thresholds, the curation

algorithms were unable to clearly delineate safety-related terms for selection, resulting in lists that were both difficult for humans to interpret and ineffective for use in safety surveillance.

Table 4. List of unigrams curated by each technique (seasonal items).

Term	Weight	First-tier piecewise Tabu search LD = 0	Second-tier piecewise Tabu search LD = 0	First-tier piecewise Tabu search LD = 1	Second-tier piecewise Tabu search LD = 1
dangerous	51,423.4	**X**		**X**	
hazard	36,845.5	**X**		**X**	
safety	27,992.4	**X**		**X**	
smelled	23,525.7	**X**		**X**	
hurt	19,006.3	**X**		**X**	
shattered	15,808.3	**X**		**X**	
caught	13,157.4	**X**		**X**	
burning	21,705.2		**X**		**X**
fire	20,286.0		**X**		**X**
cancer	17,698.4		**X**		**X**
melted	17,244.9		**X**		**X**
injured	14,263.2		**X**		**X**
danger	14,263.2		**X**		**X**
causing	13,634.4		**X**		**X**
finger	12,159.6		**X**		
careful	12,014.4		**X**		**X**

Beyond the use of fuzzy matching for term selection, we also experimented with fuzzy matching purely within our holdout set. Fuzzy matching tended to yield either similar (Levenshtein distance of 1) or worse (Levenshtein distance of 2 or beyond) performance in selecting terms, but we also tested its ability to detect reviews of interest in the holdout set. Similar to the results from term curation, however, we determined that performance was relatively similar when using a threshold of 1 but began to suffer thereafter. In Figure 5, we present a lift chart comparing the performance of the piecewise Tabu search in the seasonal items product category with exact matching (Levenshtein distance of 0) to Levenshtein distance thresholds of 1 and 2. Performance was relatively similar when comparing exact matching to the

threshold of 1. The curve incorporating fuzzy matching is slightly smoother and, in some cases, offers slightly higher performance than exact matching before tailing off around the 600th-ranked review. However, once the threshold was increased to 2, the performance of the technique was quite limited and similar to that of the sentiment analysis baseline techniques. In Table 12 of Appendix A, we present some exemplar fuzzy matches for Levenshtein distances of 1 and 2 when analyzing our seasonal items unigrams. At a threshold of 1, the fuzzy matching catches some misspellings (e.g., "hazzard" to "hazard") and similar words (e.g., "safely" to "safety") as expected but also introduces some false positives (e.g., "finer" to "finger"). At a greater threshold such as 2, false positives become more prevalent, and the matching terms are not necessarily similar in meaning (e.g., "dander" to "cancer"). Thus, we conclude that the impact of fuzzy matching is minimal when requiring close matches, and performance quickly diminishes thereafter[1].

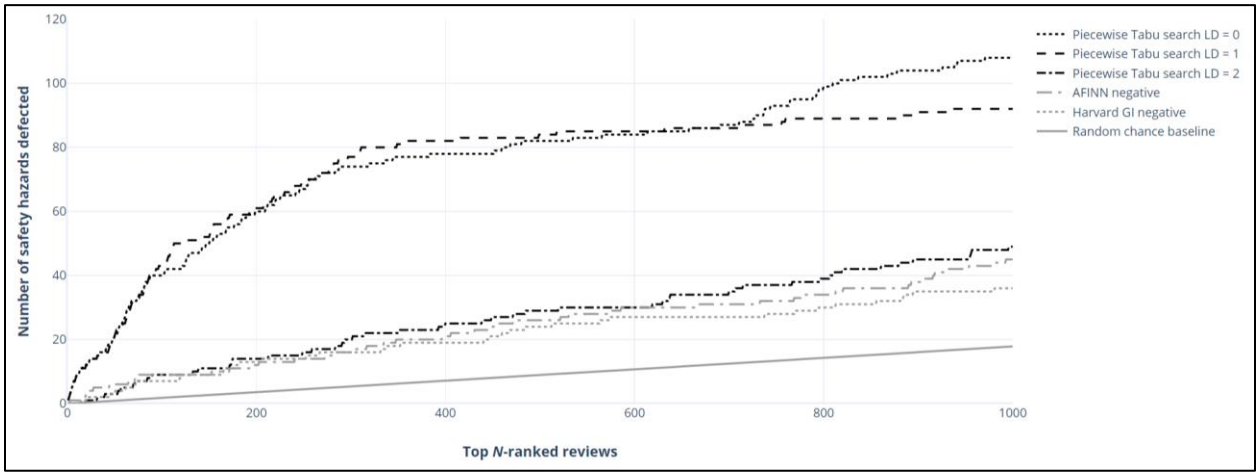

Figure 5. Lift chart comparing performance of each technique (seasonal items product category).

[1] Fuzzy matching with Levenshtein distances is a computationally intensive process for which contemporary algorithms require quadratic or near-quadratic times [5]. Future research may experiment with manual review of fuzzy matches to retain only those with close meaning, thus potentially reducing false positives and enabling shorter runtimes.

5.3 Cross-category performance

As a final stage of our analysis, we attempted to identify commonalities between smoke term lists identified in our work and those identified in past studies. If any terms are shared between these categories, then there is potential for construction of a cross-category term list that would be of use in product categories for which more finely-tuned machine-learned lists have yet to be constructed. Based on our Tables in Appendix A, we found that a total of 14 terms identified in our work matched terms that had been previously identified in other works[2]. In Table 5, we present these 14 terms together as a candidate cross-category term list. The weights that we determined for each of our terms were category-specific in our prior analyses, so to avoid biasing our analysis in favor of a particular product category, we simply assign each term an equal weight.

Table 5. Cross-category term list.

Term
burning
caught
dangerous
fire
hazard
injuries
safety
unsafe
caught fire
fire hazard
not recommend
on fire
a fire hazard
caught on fire

[2] Goldberg and Abrahams [22] also examined the OTC medicine category, so our first-tier Tabu searches on the OTC medicine product category overlap with this prior work. We exclude these terms from our analysis, as they were identified for the same product category twice as opposed to being identified in two or more distinct product categories.

12 of these terms were also included in our seasonal items smoke term lists, while just 2 were included in the OTC medicine smoke terms list. As the seasonal items product category is quite diverse, this facet may have contributed to the generality of the selected terms. Many of the terms selected in this 14-term list appear intuitively to generalize well across categories. For example, terms like “dangerous,” “hazard,” and “safety” are general safety-related terms that are not obviously related to one specific product category. Thematically, the more detailed terms are often fire-related, such as “burning,” “fire,” “caught fire,” “fire hazard,” etc. Fire hazards are common in many product categories and have previously been noted in safety surveillance work analyzing product categories such as countertop appliances [22] and dishwashers [35]. However, these terms are not limited to products with electronic components. For example, Adams et al. [4] utilized smoke terms such as “fire” and “burn” in the joint and muscle pain category, as consumers described a burning feeling associated with the use of these products. Thus, even seemingly fire-related terms may have utility across categories.

In Figure 6, we present lift charts comparing the performance of this cross-category term list to our unigram methods in each product category. In the OTC medicine category, the performance of the term list was excellent for precision, identifying 14 true safety hazards within the top 50-ranked reviews, or 0.280 precision, 0.064 recall, and 8.826 lift. Although this performance was excellent over the top 50-ranked reviews, the depth of the solution was limited, as reviews beyond this point were largely classified at the rate of random chance. Thus, the AUC for this technique was just 0.578. Interestingly, the depth of the solution was more impressive in the seasonal items product category. Within this product category, the cross-category term list did not offer the immediate precision that it did in OTC medicine, although the performance was still considerably superior to sentiment analysis or random chance. Over the top-ranking set of

reviews, the more specialized smoke term lists offered superior performance. Performance of the cross-category list was still adequate, and after the top 650-ranked reviews, performance was essentially the same as the more fine-tuned techniques. The AUC value for this approach was 0.660, which outperformed all but the unigram techniques. Although more category-specific techniques did tend to outperform the cross-category term list overall, its performance was quite competitive. This smoke term list offers a compelling heuristic for reasonably high-performing classification in product categories for which a more fine-tuned list has yet to be developed.

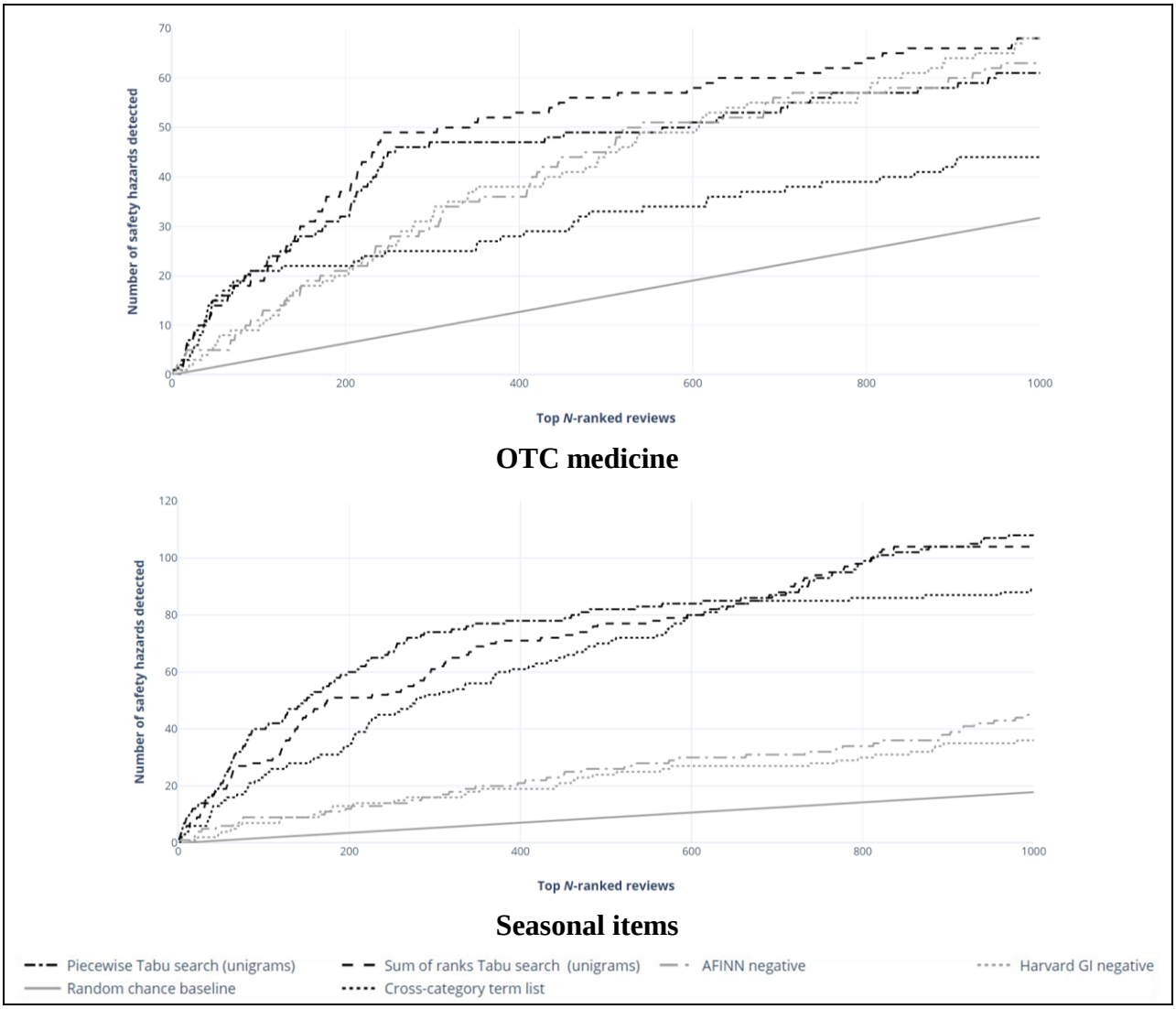

Figure 6. Lift charts comparing performance of each technique to cross-category terms.

6.0 Limitations

We note several potential limitations relevant to our study. First, our study required that humans manually tagged tens of thousands of online reviews for use in our techniques. Given our supervised learning approach, this step was necessary, but it also introduces some subjectivity as it is possible for different taggers to disagree. Despite this limitation, we aimed to validate that student taggers were reliable both internally (they agreed with one another) and externally (they agreed with a trusted authority), and we were encouraged by the substantial levels of agreement in both senses and for both datasets. Another possible limitation is that our taggers labeled an entire review in each tag rather than a smaller portion such as sentence or phrase. Tagging smaller segments of text may improve classification performance by eliminating surrounding terms that could potentially confuse classification. However, this level of specificity would require a substantial labor requirement as it would effectively require more tags over the same volume of reviews.

We also note that our analyses in this paper pertained to online reviews from two key product categories. The online reviews were obtained from two platforms, namely from Amazon.com and from the site of a Fortune 50 retailer with which we collaborated on this work. While our techniques did perform well across multiple platforms, we did not extend our work in this paper outside of online reviews to include other social media platforms, blogs, forums, etc. We believe that many of the terms generated in our techniques may be applicable to these platforms, but it is also possible that some platform-specific differences exist. Particularly for our document-based methods, alternative platforms may result in posts of different lengths and formats. Future research may explore these linguistic differences in more detail.

Online reviews as a data source are subject to several potential biases. Namely, online reviews are prone to self-selection biases because reviewers are not randomly sampled consumers; instead, they volunteer to post reviews [37]. Hu et al. [28] further add that online reviews are prone to purchasing bias – consumers purchase products that they believe they will enjoy, so reviews tend to skew positive – and under-reporting bias – consumers with extreme experiences are more likely to put forth the effort to write about their experiences than consumers with moderate experiences. These biases are important to account for in examining online reviews, but we do not believe that they are very problematic for this study. Although reviewers are not randomly sampled, we do not have reason to believe that they differ demographically from other consumers that purchased the same product. Moreover, the purchasing bias that drives online reviews to primarily be positive equally applies to consumers that do not post online. Under-reporting bias serves to accentuate the proportion of extreme reviews observed. For our application in safety hazard detection, this ensures that consumers experiencing safety hazards are likelier to write about those concerns online.

Lastly, we note that the analysis of online reviews should constitute one part of a larger effort on the part of firms to ensure the quality and safety of their products. Consumers have potentially several avenues to report hazardous products, including reporting the hazard to manufacturers directly or reporting the hazard to the appropriate regulatory agency. Similarly, firms can also test their product extensively internally to ensure appropriate quality. After products have been released to market, online reviews represent one of the most voluminous and rapidly updating manners in which consumers can explain their experiences with a product, and thus they ought to be monitored substantially by stakeholders, but this monitoring should not be to the exclusion of other key measures.

7.0 Conclusion, future work, and implications

In this study, we examine several novel methodological enhancements for safety surveillance of online media. We augment prior methodologies that utilize smoke terms, or terms specifically tuned to hazard-related discussions within a particular product category. Goldberg and Abrahams [22] suggested the use of a Tabu search heuristic for providing high-precision smoke term lists. We consider cases in which the depth of the solution is also valuable and provide a series of experiments on a variety of methods. We found that piecewise Tabu search smoke term lists offer the same benefits of precision as a single tier of smoke terms, but the addition of a second tier improves the depth of the solution once matches to the initial list are exhausted. We also experimented with a sum of ranks Tabu search objective, which provides slightly different sets of terms that offered competitive performance. We found that fuzzy matching did not substantially improve performance either in the curation step or in the holdout set. With a restrictive threshold for fuzzy matching, results were quite similar to exact matching, and once the threshold was increased, performance diminished. Finally, we compared the smoke terms identified in this work to smoke terms identified in past works, devising a cross-category smoke term list from the terms that overlapped. Although this list did not perform as well as category-specific smoke term lists, its performance was largely competitive, suggesting that it is a viable option for quick analysis in product categories for which a fine-tuned smoke term list has yet to be developed.

This study has implications for researchers, industry, and regulators alike. One major implication from this study is the potential for smoke term lists that improve the depth of solutions using piecewise or sum of ranks Tabu search functions. For instances in which every true positive is especially valuable or in which an organization has a great deal of resources to

devote to safety surveillance efforts, these improvements are invaluable. As opposed to prior techniques that perform well only in a top-ranking set of reviews, these solutions span a larger portion of a corpus. Research may continue to experiment with solutions that improve the depth of solutions in addition to preserving precision.

Our work also implies the potential for our smoke term lists to be applied to safety monitoring within multiple product categories. Our smoke term lists specifically curated for the OTC medicine and seasonal items product categories offer excellent performance both over a top-ranking set of reviews and beyond. Users must choose the compromise between rapid prioritization and depth that best fits their application, but out smoke term lists ensure that they have a variety of options. Beyond these two product categories, we also provide a cross-category smoke term list that may be applicable across a wide variety of circumstances. In our experiments, we found that the list did not perform as well as a more finely-tuned category-specific list, but performance was still a substantial improvement over baseline methods. Future research ought to experiment with the development of these cross-category lists and examine their performance across a variety of potential settings.

References

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

[2] 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, Decision Support Systems, 55(4) (2013) 871-882.

[3] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decision Support Systems, 54(1) (2012) 87-97.

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

[5] A. Backurs, P. Indyk, Edit distance cannot be computed in strongly subquadratic time (unless SETH is false), in: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, (ACM, 2015), pp. 51-58.

[6] T. Bao, T.-l.S. Chang, Why Amazon uses both the New York Times Best Seller List and customer reviews: An empirical study of multiplier effects on product sales from multiple earned media, Decision Support Systems, 67 (2014) 1-8.

[7] J. Berger, A.T. Sorensen, S.J. Rasmussen, Positive effects of negative publicity: When negative reviews increase sales, Marketing Science, 29(5) (2010) 815-827.

[8] BrightLocal, Local Consumer Review Survey, in, (2016).

[9] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: Online book reviews, Journal of Marketing Research, 43(3) (2006) 345-354.

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

[11] S. Debortoli, O. Müller, I.A. Junglas, J. vom Brocke, Text mining for information systems researchers: An annotated topic modeling tutorial, Communications of the AIS, 39 (2016) 7.

[12] I.R. Edwards, M. Lindquist, Social media and networks in pharmacovigilance, in, (Springer, 2011).

[13] W. Fan, M.D. Gordon, The power of social media analytics, Communications of the ACM, 57(6) (2014) 74-81.

[14] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decision Support Systems, 40(2) (2005) 213-233.

[15] W. Fan, L. Wallace, S. Rich, Z. Zhang, Tapping the power of text mining, Communications of the ACM, 49(9) (2006) 76-82.

[16] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, 27(8) (2006) 861-874.

[17] J. Fingas, Amazon shopping test recommends products based on your likes, in: Engadget, (2018).

[18] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin, 76(5) (1971) 378.

[19] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3(Mar) (2003) 1289-1305.

[20] F. Glover, Future paths for integer programming and links to artificial intelligence, Computers & Operations Research, 13(5) (1986) 533-549.

[21] D. Goldberg, N. Zaman, Text analytics for employee dissatisfaction in human resources management, in: 24th Americas Conference on Information Systems, (2018).

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

[23] R. Gruss, Text analytics for customer engagement in social media, in: Business Information Technology, (Virginia Tech, 2018).

[24] L. Hazell, S.A. Shakir, Under-reporting of adverse drug reactions, Drug Safety, 29(5) (2006) 385-396.

[25] S.R. Hiltz, M. Turoff, Structuring computer-mediated communication systems to avoid information overload, Communications of the ACM, 28(7) (1985) 680-689.

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

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

[28] N. Hu, J. Zhang, P.A. Pavlou, Overcoming the J-shaped distribution of product reviews, Communications of the ACM, 52(10) (2009) 144-147.

[29] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering, 17(3) (2005) 299-310.

[30] K.-L. Huang, C.-W. Kuo, M.-L. Lu, Wholesale price rebate vs. capacity expansion: The optimal strategy for seasonal products in a supply chain, European Journal of Operational Research, 234(1) (2014) 77-85.

[31] D. Joo, T. Hong, I. Han, The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors, Expert Systems with Applications, 25(1) (2003) 69-75.

[32] E.F. Kelly, P.J. Stone, Computer recognition of English word senses, (North-Holland, 1975).

[33] G. Kulkarni, P. Kannan, W. Moe, Using online search data to forecast new product sales, Decision Support Systems, 52(3) (2012) 604-611.

[34] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics, (1977) 159-174.

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

[36] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady, (1966), pp. 707-710.

[37] X. Li, L.M. Hitt, Self-selection and information role of online product reviews, Information Systems Research, 19(4) (2008) 456-474.

[38] Y.-M. Li, T.-Y. Li, Deriving market intelligence from microblogs, Decision Support Systems, 55(1) (2013) 206-217.

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

[40] E. Lopez-Gonzalez, M.T. Herdeiro, A. Figueiras, Determinants of under-reporting of adverse drug reactions, Drug Safety, 32(1) (2009) 19-31.

[41] A. Marucheck, N. Greis, C. Mena, L. Cai, Product safety and security in the global supply chain: Issues, challenges and research opportunities, Journal of Operations Management, 29(7-8) (2011) 707-720.

[42] J. McAuley, R. Pandey, J. Leskovec, Inferring networks of substitutable and complementary products, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM, 2015), pp. 785-794.

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

[44] F.Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, in: Proceedings of the 1st Workshop on Making Sense of Microposts, (2011), pp. 93-98.

[45] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, (2011).

[46] Z. Qiao, G.A. Wang, M. Zhou, W. Fan, The impact of customer reviews on product innovation: empirical evidence in mobile apps, in: Analytics and Data Science, (Springer, 2018), pp. 95-110.

[47] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, (Association for Computational Linguistics, 2009), pp. 147-155.

[48] E. Riloff, W. Lehnert, Information extraction as a basis for high-precision text classification, ACM Transactions on Information Systems, 12(3) (1994) 296-333.

[49] N.G. Rupp, The attributes of a costly recall: Evidence from the automotive industry, Review of Industrial Organization, 25(1) (2004) 21-44.

[50] S. Saluja, S. Woolhandler, D.U. Himmelstein, D. Bor, D. McCormick, Unsafe drugs were prescribed more than one hundred million times in the United States before being recalled, International Journal of Health Services, 46(3) (2016) 523-530.

[51] S. Seo, S.S. Jang, L. Miao, B. Almanza, C. Behnke, The impact of food safety events on the value of food-related firms: An event study approach, International Journal of Hospitality Management, 33(2013) 153-165.

[52] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research, 2(Nov) (2001) 45-66.

[53] J. Vörös, On the risk-based aggregate planning for seasonal products, International Journal of Production Economics, 59(1-3) (1999) 195-201.

[54] J.D. Weidenhamer, Lead contamination of inexpensive seasonal and holiday products, Science of the Total Environment, 407(7) (2009) 2447-2450.

[55] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decision Support Systems, 90 (2016) 23-32.

[56] J. Woodcock, A difficult balance—pain management, drug safety, and the FDA, New England Journal of Medicine, 361(22) (2009) 2105-2107.

[57] K. Xu, S.S. Liao, J. Li, Y. Song, Mining comparative opinions from customer reviews for competitive intelligence, Decision Support Systems, 50(4) (2011) 743-754.

[58] B. Yu, Z.-b. Xu, A comparative study for content-based dynamic spam classification using four machine learning algorithms, Knowledge-Based Systems, 21(4) (2008) 355-362.

[59] H. Yuan, R.Y. Lau, W. Xu, The determinants of crowdfunding success: A semantic text analytics approach, Decision Support Systems, 91 (2016) 67-76.

[60] X. Zhao, Y. Li, B.B. Flynn, The financial impact of product recall announcements in China, International Journal of Production Economics, 142(1) (2013) 115-123.

[61] X. Zu, L.D. Fredendall, T.J. Douglas, The evolving theory of quality management: The role of Six Sigma, Journal of Operations Management, 26(5) (2008) 630-650.

Appendix A: Supplementary material

Table 6. List of unigrams curated by each technique (OTC medicine).

Term	Weight	First-tier piecewise Tabu search	Second-tier piecewise Tabu search	Sum of ranks Tabu search	Prior research
capful	4,966.7	**X**		**X**	[22]
caution	4,626.5	**X**		**X**	[22]
liver	4,507.7	**X**		**X**	[22]
potentially	3,990.6	**X**		**X**	[22]
surrounding	3,990.6	**X**		**X**	[22]
caused	3,567.8	**X**		**X**	[22]
catastrophic	3,511.5	**X**		**X**	[22]
overdosed	3,511.5	**X**			[22]
chills	3,511.5	**X**		**X**	[22]
monster	3,511.5	**X**			[22]
heaving	3,511.5	**X**		**X**	[22]
toll	4,966.7		**X**	**X**	
damage	3,619.5		**X**		
misled	3,511.5		**X**		
unsafe	3,511.5		**X**	**X**	[3, 43]
theories	3,511.5		**X**	**X**	
induce	3,511.5			**X**	
horribly	3,511.5			**X**	
counterproductive	3,511.5			**X**	

Table 7. List of unigrams curated by each technique (seasonal items).

Term	Weight	First-tier piecewise Tabu search	Second-tier piecewise Tabu search	Sum of ranks Tabu search	Prior research
dangerous	51,423.4	**X**		**X**	[3, 4, 22, 43]
hazard	36,845.5	**X**		**X**	[22, 43]
safety	27,992.4	**X**		**X**	[22, 43]
smelled	23,525.7	**X**		**X**	
hurt	19,006.3	**X**			
shattered	15,808.3	**X**		**X**	
caught	13,157.4	**X**			[4, 22, 35]
burning	21,705.2		**X**	**X**	[35]
fire	20,286.0		**X**	**X**	[4, 22, 55]
cancer	17,698.4		**X**	**X**	
melted	17,244.9		**X**	**X**	
injured	14,263.2		**X**	**X**	
danger	14,263.2		**X**		
causing	13,634.4		**X**	**X**	
finger	12,159.6		**X**	**X**	
careful	12,014.4		**X**	**X**	
smell	20,427.1			**X**	
injuries	16,793.8			**X**	[55]
flew	15,452.8			**X**	
explode	13,118.4			**X**	

Table 8. List of bigrams curated by each technique (OTC medicine).

Term	Weight	First-tier piecewise Tabu search	Second-tier piecewise Tabu search	Sum of ranks Tabu search	Prior research
stomach pain	5,185.0	**X**		**X**	[22]
caution if	4,966.7	**X**		**X**	[22]
by afternoon	4,966.7	**X**		**X**	[22]
it caused	4,966.7	**X**		**X**	[22]
with caution	4,612.6	**X**		**X**	[22]
periods of	3,990.6	**X**		**X**	[22]
that evening	3,990.6	**X**		**X**	[22]
severely burnt	3,511.5	**X**			[22]
your belly	3,511.5	**X**			[22]
cold sweats	4,966.7		**X**	**X**	
reaction to	4,165.1		**X**	**X**	
also noticed	3,990.6		**X**	**X**	
and cramping	3,990.6		**X**	**X**	
cause drowsiness	3,990.6		**X**	**X**	
after burn	3,511.5		**X**		
have caused	4,966.7			**X**	
not recommend	4,507.7			**X**	[4]
hurt your	3,990.6			**X**	
cause rectal	3,511.5			**X**	

Table 9. List of bigrams curated by each technique (seasonal items).

Term	Weight	First-tier piecewise Tabu search	Second-tier piecewise Tabu search	Sum of ranks Tabu search	Prior research
safety hazard	29,314.7	**X**		**X**	
fire hazard	27,562.3	**X**		**X**	[22]
very dangerous	26,848.2	**X**		**X**	
burning smell	24,149.4	**X**		**X**	
on fire	23,347.1	**X**		**X**	[22]
had melted	20,568.9	**X**			
black smoke	17,698.4	**X**		**X**	
million pieces	17,695.0	**X**		**X**	
started smoking	16,680.1	**X**		**X**	
caught fire	15,727.3	**X**		**X**	[22]
a safety	14,049.9	**X**		**X**	
the safety	13,690.5	**X**		**X**	
be careful	13,469.1	**X**		**X**	
fire and	12,603.2	**X**		**X**	
like poison	11,874.5	**X**			
be recalled	20,568.9		**X**	**X**	
caught on	18,416.8		**X**		
shattered into	17,698.4		**X**	**X**	
caked on	16,793.8		**X**		
dangerous product	16,793.8		**X**	**X**	
flew out	16,793.8		**X**	**X**	
fly out	16,793.8		**X**	**X**	
injured while	16,793.8		**X**	**X**	
pit rusted	16,793.8		**X**	**X**	
recalled by	16,793.8		**X**	**X**	
leaks gas	13,579.2		**X**		
but dangerous	16,793.8			**X**	
dangerous and	16,793.8			**X**	
smells like	14,263.2			**X**	
fire i	13,839.3			**X**	
the fire	13,798.4			**X**	
called char	13,579.2			**X**	
of burning	13,579.2			**X**	

Table 10. List of trigrams curated by each technique (OTC medicine).

Term	Weight	First-tier piecewise Tabu search	Second-tier piecewise Tabu search	Sum of ranks Tabu search	Prior research
my stomach is	5,212.9	**X**		**X**	[22]
pain and cramping	4,966.7	**X**		**X**	[22]
wouldn't use it	4,966.7	**X**		**X**	[22]
going to faint	4,966.7	**X**		**X**	[22]
had this problem	4,966.7	**X**		**X**	[22]
and very bad	4,966.7	**X**		**X**	[22]
with caution if	4,966.7	**X**		**X**	[22]
it with caution	4,966.7	**X**		**X**	[22]
this product contains	3,990.6	**X**			[22]
sweats and flu	3,511.5	**X**		**X**	[22]
all day now	3,511.5	**X**		**X**	[22]
morning and still	3,511.5	**X**		**X**	[22]
actually made me	4,966.7		**X**	**X**	
afternoon i had	4,966.7		**X**		
next day i	3,990.6		**X**	**X**	
almost burn with	3,511.5		**X**	**X**	
burns your throat	3,511.5		**X**	**X**	
had cold sweats	3,511.5		**X**	**X**	
throwing up everywhere	3,511.5		**X**	**X**	
was spoiled or	3,511.5		**X**		
by afternoon i	4,966.7			**X**	
problem with it	4,966.7			**X**	

Table 11. List of trigrams curated by each technique (seasonal items).

Term	Weight	First-tier piecewise Tabu search	Second-tier piecewise Tabu search	Sum of ranks Tabu search	Prior research
a fire hazard	23,751.9	**X**		**X**	[22]
caught on fire	22,271.6	**X**		**X**	[22]
should be recalled	20,568.9	**X**		**X**	
a million pieces	17,695.0	**X**		**X**	
a burning smell	16,793.8	**X**		**X**	
a safety hazard	16,793.8	**X**		**X**	
be recalled by	16,793.8	**X**		**X**	
dangerous to use	16,793.8	**X**		**X**	
fire hazard but	16,793.8	**X**		**X**	
fire hazard this	16,793.8	**X**		**X**	
to get hurt	16,793.8	**X**			
a very dangerous	13,579.2	**X**		**X**	
on fire i	21,141.9		**X**	**X**	
so called safety	20,568.9		**X**	**X**	
could have burned	16,793.8		**X**	**X**	
it smokes like	16,793.8		**X**		
it to explode	16,793.8		**X**		
be a safety	13,579.2		**X**	**X**	
glass on the	13,579.2		**X**	**X**	
to blow up	13,579.2		**X**	**X**	
burned me or	11,874.5		**X**		
on fire i	21,141.9			**X**	
so called safety	20,568.9			**X**	
could have burned	16,793.8			**X**	
it smokes like	16,793.8			**X**	
it to explode	16,793.8			**X**	

Table 12. Exemplar fuzzy term matches (seasonal items unigrams).

Fuzzy matching threshold	**Exemplar matches**
Levenshtein distance = 1	“dangerousl” to “dangerous” “finer” to “finger” “hazzard” to “hazard” “safely” to “safety” “swelled” to “smelled”
Levenshtein distance = 2	“calling to “causing” “cut” to “hurt” “dander” to “cancer” “dangerously” to “dangerous” “hard” to “hazard”

Conclusion

This dissertation presents a collection of studies examining different methods and use cases for extracting business intelligence from online reviews. Each study utilizes text analytics, using distinctive machine-learned smoke terms to identify reviews of interest. The first study applies these analytics in the domain of safety surveillance, proposing a heuristic for choosing smoke terms that efficiently prioritize online reviews that pertain to safety hazards. This automation offers substantial time savings relative to manually parsing each review. The second study extends these analytics to product innovation. Building upon the well-accepted attribute mapping framework [7, 8], we aim to identify reviews that mention particular attributes or features of products that differentiate them from the competition, whether positively or negatively. We develop smoke terms to detect these attributes, and we validate that product managers find the retrieved reviews useful to their innovation processes. The third study addresses safety surveillance from a standpoint that incorporates the total yield of safety hazards found in addition to efficiency. We perform computational experiments using a variety of techniques ranging from alternative heuristic objectives to fuzzy matching. Finally, we construct cross-category smoke term lists and assess their performance compared to our more finely-tuned techniques.

There are many avenues for future research streams extending from these studies. The third study explores the potential for the construction of cross-category smoke term lists by identifying overlap in smoke terms between that study and past studies. We found that the proposed smoke term list is applicable across multiple categories, although category-specific smoke term lists outperformed it as expected. Future works could apply this smoke term list in other settings or could experiment with other means of constructing a cross-category list.

Furthermore, the use of a hybrid of several methods may be compelling. The first study found that consumers' star ratings were useful as a further indicator in safety surveillance, and each study in this dissertation has found that sentiment analyses outperform random chance. A holistic model that incorporates each of these components would have great potential.

The third study examined the use of fuzzy matching to improve the depth of the solutions offered by text analytic techniques. As the fuzzy matching appeared to introduce many false positives in addition to true positives (especially at more lenient matching thresholds), this technique did not substantially improve performance. An alternative technique to implement in future research is to use automated spellchecking technology to account for misspellings of smoke terms. Such a technique would ideally exclude fuzzy matches of similarly spelled but semantically disparate terms, thus improving performance.

A further consideration for future research involves rule induction for these text analytics. Each study in this dissertation generates smoke terms that are each associated with linear weights (see Chapter 1, Appendix A for more details). Future research may consider interactions between smoke terms such that they behave nonlinearly. For example, perhaps the combination of "safety" and "hazard" suggests that the smoke score for the associated review should be greater than the sum of the two terms. Alternatively, perhaps "safety" should generally be taken to refer to safety hazards, but this rule should be ignored if it is used as a part of the bigram "safety feature." Finally, research may also experiment with terms that exclude a review from classification as a safety hazard, such as terms like "best" or "excellent."

On a higher level, each study in this dissertation is built upon supervised machine learning algorithms trained with human-tagged data. The process of manually tagging this data can be quite time-consuming and potentially expensive, and the insights derived from the

machine learning are only valuable to the extent that the underlying data is appropriately labeled. The literature acknowledges the complex nature of collaborative tagging systems (e.g., [1]), but it has largely lacked a true assessment of the reliability of the tags that these systems generate. Content analysis literature proposes many metrics to assess the accuracy of tagging in post-processing [2, 4, 5], but little guidance is available for modern collaborative tagging platforms in which many taggers may be tagging simultaneously, each being paid as they tag. Today, many studies depend on outsourced tagging through online systems such as Amazon Mechanical Turk or Amazon SageMaker Ground Truth [9], although the use of these platforms is controversial [3, 6]. This configuration raises issues of reliability on a tagger-level, as unreliable taggers' data may need to be discarded, and real-time detection of unreliable work eases the financial burden of the tagging project on organizations.

These challenges present several possible directions for future research. First, many of the current reliability metrics are typically applied in post-processing rather than in real-time, as they may rely on complete datasets without missing values [4, 5]. Development of real-time measures could allow researchers or practitioners to make vital adjustments before work is wasted. Second, the metrics assess the reliability of the entire dataset of tags as opposed to the reliability of specific tags or taggers. Constructing metrics that apply on a finer level could allow for poor tagging to be corrected rapidly or, if necessary, for "rogue" taggers to be dismissed. Third, these metrics give no consideration to the amount of effort required in tagging projects. As these projects have become massive with the advent of Amazon Mechanical Turk and other tools, minimizing the amount of unreliable tagging is an essential element of cost control. Finally, an adaptive tagging approach may be a viable method for this cost control consideration. Rather than performing all machine learning analyses after tagging is completed, these analyses

could be performed in real-time, suggesting smoke terms while taggers are still working. Then, rather than taggers continuing to tag random reviews, their tasks could prioritize reviews that contain the already detected smoke terms. The text analytics techniques would further refine term selection and suggestions for reviews to tag as more data was generated.

Online reviews represent a compelling opportunity for firms to learn more about both their products and their consumers, deriving actionable business intelligence from consumer feedback. Firms that capitalize upon these insights have the opportunity to more quickly adapt to market conditions, replacing or updating underperforming or potentially hazardous products to improve their competitive position. In this dissertation, we present a series of studies that propose text analytic techniques for automating the extraction of vital insights from these reviews. Advances in text analytics will continue to advance firms' surveillance efforts and thereby improve product quality.

References

[1] H. Halpin, V. Robu, H. Shepherd, The complex dynamics of collaborative tagging, in: Proceedings of the 16th International Conference on World Wide Web, (ACM, 2007), pp. 211-220.

[2] A.F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data, Communication Methods and Measures, 1(1) (2007) 77-89.

[3] R. Jia, Z.R. Steelman, B.H. Reich, Using Mechanical Turk data in IS research: Risks, rewards, and recommendations, Communications of the AIS, 41(2017) 14.

[4] K. Krippendorff, Reliability in content analysis, Human Communication Research, 30(3) (2004) 411-433.

[5] K. Krippendorff, Agreement and information in the reliability of coding, Communication Methods and Measures, 5(2) (2011) 93-112.

[6] P.B. Lowry, J. D'Arcy, B. Hammer, G.D. Moody, "Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels, Journal of Strategic Information Systems, 25(3) (2016) 232-240.

[7] I.C. MacMillan, R.G. McGrath, Discover your products' hidden potential, Harvard Business Review, 74(3) (1996) 58-73.

[8] I.C. MacMillan, R.G. McGrath, Discovering new points of differentiation, Harvard Business Review, 75(1997) 133-145.

[9] Z.R. Steelman, B.I. Hammer, M. Limayem, Data Collection in the Digital Age: Innovative Alternatives to Student Samples, MIS Quarterly, 38(2) (2014).