

**Online Review Analytics: New Methods for discovering
Key Product Quality and Service Concerns**

Nohel Zaman

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

Cliff T. Ragsdale

Roberta S. Russell

Michelle M.H. Seref

G. Alan Wang

May 10, 2019
Blacksburg, VA

Keywords: online reviews, text analytics, risk assessment, hospitals, service
quality.

© Nohel Zaman

Online Review Analytics: New Methods for discovering Key Product Quality and Service Concerns

Nohel Zaman

Abstract (academic)

The purpose of this dissertation intends to discover as well as categorize safety concern reports in online reviews by using key terms prevalent in sub-categories of safety concerns. This dissertation extends the literature of semi-automatic text classification methodology in monitoring and classifying product quality and service concerns. We develop various text classification methods for finding key concerns across a diverse set of product and service categories. Additionally, we generalize our results by testing the performance of our methodologies on online reviews collected from two different data sources (Amazon product reviews and Facebook hospital service reviews). Stakeholders such as product designers and safety regulators can use the semi-automatic classification procedure to subcategorize safety concerns by injury type and narrative type (Chapter 1). We enhance the text classification approach by proposing a Risk Assessment Model for quality management (QM) professionals, safety regulators, and product designers to allow them to estimate overall risk level of specific products by analyzing consumer-generated content in online reviews (Chapter 2). Monitoring and prioritizing the hazard risk levels of products will help the stakeholders to make appropriate actions on mitigating the risk of product safety. Lastly, the text classification approach discovers and ranks aspects of services that predict overall user satisfaction (Chapter 3). The key service terms are beneficial for healthcare providers to rapidly trace specific service concerns for improving the hospital services.

Online Review Analytics: New Methods for discovering Key Product Quality and Service Concerns

Nohel Zaman

Abstract (public)

This dissertation extends past studies by examining safety surveillance of online reviews. We examine online reviews reporting specific categories of safety concerns and contrast them with reviews not reporting these specific safety concerns. Business and regulators are benefited in detecting, categorizing, and prioritizing safety concerns across product categories. We use key terms prevalent in domain-related safety concerns for granular analysis of consumer reviews. Secondly, beyond utilizing the key terms to discover specific hazard incidents, safety regulators and manufacturers may use the extended risk assessment framework to estimate the risk severity, risk likelihood, and overall risk level of a specific product. The model could be useful for product safety practitioners in product risk identification and mitigation. Finally, this dissertation identifies the aspects of service quality concerns present in online hospital reviews. This study uses text analytics method by using key terms to detect these specific service concerns and hence determine primary rationales for patient feedback on hospital services. Managerially, this information helps to prioritize the areas in greatest need of improvement of hospital services. Additionally, generating key terms for a particular service attribute aids health care policy makers and providers in rapidly monitoring specific concerns and adjusting policies or resources to better serve patient

Acknowledgements

First of all, this work of my dissertation would not have been possible without the support of my peers and undergraduate students of Business Information Technology (BIT) at Virginia Tech. Especially, my peer, David Michael Goldberg had a significant contribution in proof-reading the content of all my chapters. Additionally, he was helpful in running the unsupervised learning technique, LDA on safety concern reports (Chapter 1) and also participated in multiple tagging assignments. I am thankful to Dr. Richard Gruss for developing and operating the PamTag collaborative tagging system which was essential for the data coding of my dissertation studies. Additionally, I am glad work with all my peers during my PhD program. For instance, my peer at the BIT department, Qianzhou Du had put a lot of effort in extracting the patient reviews from the Facebook pages of hospitals for my Chapter 2 study. I would also like to thank the five product safety professionals who volunteered time to complete my product risk assessment survey, in Chapter 2.

An undergraduate student from the BIT department, Arsham Zamani invested a substantial amount of his time in tagging reviews for multiple projects (i.e. 6 different tagging assignments: Event, Hazard, and Injury categories on two different datasets). There are also multiple undergraduates at Virginia Tech whose datasets I leveraged in Chapters 1 and 2: Laura Sierra, Taylor Dorn, Nick Carroll, Matthew Winkler, Vaibhav Mummalaneni, Brandon Kim, Melissa Nipper, Darren Law, Aaron Kim, Sana Rauf, and Leigh Midkiff.

I received a wide range of professional feedback from all the members of my Dissertation Committee when searching for my job in a faculty and also gained a lot of knowledge in quantitative research from them. I am blessed and also fortunate to have Dr. Alan S. Abrahams

as the chairperson of my committee. As my guru as well as my advisor, he has guided me with my research and teaching through this whole journey of my PhD. program. I saw him as a real-life example what a researcher (and a patient person) must be when pursuing challenging goals in life. It's very hard for me to write or express in words how much effort, feedback and sleepless nights my advisor had to sacrifice in order for me to prosper in my academic life!

Last but not least, I am very fortunate to receive the enthusiasm and encouragement from the members of my family in pursuing my academic goals and completing my dissertation work. I appreciate my parents' guidance and prayers that are always with me. Predominantly, I am thankful to my beloved and my caring wife, Rawfin who provides me with endless inspiration and also helping me with formatting my dissertation. Also, I am blessed to have a wonderful son, Nayef whose unlimited joy and smile boost my energy every day to go forward.

Note

I contributed a portion of the dissertation work to the following publication:

Zaman, N., Goldberg, D.M., Gruss, R.J., Ehsani, J.P., and Abrahams, A.S. “Using text analytics to classify safety hazard reports in online reviews.” 2018. *2018 INFORMS Workshop on Data Science*. Phoenix, AZ.

Table of Contents

INTRODUCTION	X
1. PRIOR WORK AND RESEARCH GAPS.....	X
1.1. ONLINE REVIEW ANALYTICS.....	XI
1.2. TEXT MINING AND TEXT CLASSIFICATION.....	XII
2. DEFINITIONS	XIII
2.1. SAFETY CONCERN SUBCATEGORIES (CHAPTER 1)	XIII
2.2. RISK ASSESSMENT MODEL (CHAPTER 2).....	XIV
2.3. HOSPITAL SERVICE CODES (CHAPTER 3).....	XV
3. TEXT-ANALYTICS-BASED METHODS FOR DISCOVERING KEY PRODUCT QUALITY AND SERVICE CONCERNS IN ONLINE REVIEWS	XV
REFERENCES.....	XVIII
CHAPTER 1. USING TEXT ANALYTICS TO SUB-CLASSIFY SAFETY-CONCERN MENTIONS IN ONLINE REVIEWS	1
ABSTRACT.....	1
1. INTRODUCTION	2
2. LITERATURE REVIEW	4
2.1. STRUCTURED DATA SOURCES.....	4
2.2. ONLINE REVIEWS.....	6
2.3. SENTIMENT ANALYSIS AND TEXT CLASSIFICATION	6
2.4. RESEARCH CONTRIBUTION AND RESEARCH GAP.....	7
3. METHODOLOGY	10
3.1. DATA SOURCE AND DATA CODING.....	10
3.2. ASSESSMENT OF SMOKE WORDS	19
4. RESULTS AND DISCUSSION	20
4.1. EXPLORATORY STUDY.....	20
4.2. SAFETY CONCERN CATEGORY-SPECIFIC ANALYSIS.....	24
4.3. GENERALIZABILITY OF CATEGORY-SPECIFIC SMOKE N-GRAMS	30
5. CONCLUSION	37
6. LIMITATIONS	37
7. IMPLICATIONS.....	38
7.1. IMPLICATION FOR PRACTICE	38
7.2. IMPLICATION FOR RESEARCH	39
8. FUTURE WORK.....	41
REFERENCES.....	42
APPENDIX A. TAGGING PROTOCOL OF THE CODE VALUES FOR EACH EVENT, HAZARD, AND INJURY CODE CATEGORY	46
APPENDIX B. SMOKE WORD LISTS FOR EACH CODE VALUE UNDER EVENT, HAZARD, AND INJURY CATEGORIES.....	60

APPENDIX C. NUMBER OF TRUE POSITIVES IN TOP-SCORED AND BOTTOM-SCORED SENTENCES, FOR EACH CODE VALUE OF EVENT, HAZARD, AND INJURY CATEGORIES.....	64
APPENDIX D. COMPARISON OF THE SMOKE SCORING METHODS WITH THE CONVENTIONAL SENTIMENT METHOD, AFINN, IN DETECTING THE CODE VALUES ACROSS THE TOP N-RANKED SENTENCES FOR EACH CODE VALUE IN EVENT, HAZARD, AND INJURY CATEGORIES.....	68
APPENDIX E. ASSESSMENT OF SMOKE SCORING METHODS IN DISCOVERING CODE VALUES OF EVENT, HAZARD, AND INJURY CATEGORIES ACROSS DIFFERENT DATASETS	74
CHAPTER 2. A SEMI-AUTOMATED RISK ASSESSMENT METHOD FOR CONSUMER PRODUCTS .	76
ABSTRACT.....	76
1. INTRODUCTION.....	76
2. LITERATURE REVIEW	82
2.1. STRUCTURED DATA	82
2.1.1. <i>Consumer safety complaints dataset (Saferproducts.gov)</i>	83
2.1.2. <i>Hospitalization dataset (NEISS)</i>	84
2.1.3. <i>Product utilization dataset (Nielsen)</i>	84
2.1.4. <i>GlobalRecalls portal</i>	85
2.2. ONLINE REVIEWS.....	86
2.3. TEXT CLASSIFICATION	87
3. RESEARCH CONTRIBUTION AND GAP.....	89
4. METHODOLOGY.....	93
4.1. DATA COLLECTION.....	93
4.2. DATA CODING AND PROCESSING.....	93
4.3. SENTENCE-LEVEL TAGGING PROCEDURE.....	95
4.4. GENERATION OF SMOKE WORDS.....	96
4.5. RISK SCORING PROCEDURE.....	97
4.5.1. <i>Estimating Hazard Likelihood for a product</i>	99
4.5.2. <i>Estimating Hazard Severity for a Product</i>	99
4.5.3. <i>Estimating Severity for a Hazard Type</i>	100
4.6. PROCEDURE FOR EVALUATING THE RISK ASSESSMENT MODEL	103
4.7. RISK MODEL ASSESSMENT WITH THE COMPUTER VERSUS HUMAN EXPERTS	105
5. RESULTS AND DISCUSSION	106
5.1. SCORING PRODUCTS	106
5.2. SEVERITY WEIGHTS FOR HAZARD CODES	110
5.3. POSITIONING PRODUCTS ON THE EU RISK MATRIX	111
5.4. RISK MODEL EVALUATION RESULTS	118
6. LIMITATIONS AND FUTURE WORK.....	120
6.1. SAMPLE SIZE AND SELECTION BIASES	120
6.2. SURVEY PARTICIPATION	121
6.3. ACCURACY IN RISK MODEL CALIBRATION	121
6.4. OBJECTIVE STATISTICAL ASSESSMENT OF THE RISK MODEL	122
6.5. MITIGATING SELECTION BIASES.....	123
6.6. ESTIMATING PRODUCT PENETRATION MORE ACCURATELY.....	123
6.7. NON-REPRESENTATIVENESS OF THE DATASET	124
6.8. FURTHER CPSC STRATEGIC GOALS.....	125

6.9. CLASSIFICATION OF INJURY NARRATIVES	126
7. CONCLUSION	127
8. IMPLICATIONS FOR RESEARCH AND PRACTICE.....	127
REFERENCES.....	130
APPENDIX A: SURVEY INSTRUMENT (SAMPLE)	134
APPENDIX B. HUMAN EXPERTS' RISK SCORES AND RATIONALES OF VARIOUS PRODUCTS ...	140
CHAPTER 3. FACEBOOK HOSPITAL REVIEWS: AUTOMATED SERVICE QUALITY DETECTION AND RELATIONSHIPS WITH PATIENT SATISFACTION	151
ABSTRACT.....	151
1. INTRODUCTION	151
2. LITERATURE REVIEW	153
2.1. ONLINE RATINGS AND PATIENT EXPERIENCES.....	153
2.2. TEXT AND SENTIMENT ANALYSIS	155
2.3. RESEARCH GAP	157
3. RESEARCH QUESTIONS AND CONTRIBUTIONS	162
4. METHODOLOGY	163
4.1. EXPLORATORY ANALYSIS	165
4.2. DATA CODING.....	166
4.3. REGRESSION ANALYSIS.....	167
4.4. SMOKE TERM GENERATION	168
4.5. ASSESSING SMOKE TERMS.....	169
5. RESULTS AND DISCUSSION	169
5.1. EXPLORATORY ANALYSIS	169
5.2. TAGGING ANALYSIS	170
5.3. REGRESSION ANALYSIS.....	174
5.4. TEXT ANALYTIC ANALYSIS.....	177
6. CONCLUSIONS.....	184
7. LIMITATIONS AND FUTURE WORK	184
REFERENCES.....	187
APPENDIX A: TAGGING PROTOCOL	191
APPENDIX B: INTER-RATER RELIABILITY METRICS.....	194
APPENDIX C: MOST PREVALENT TERMS FOR EACH SERVICE ATTRIBUTE VALUE.....	196
CONCLUSION, IMPLICATIONS, AND FUTURE WORK	202

Introduction

Due to the continuous growth of user-generated content (UGC) in social media, companies now have valuable new resources at their disposal for monitoring consumer feedback on product quality. When posting reviews, social media posts, forum posts, or other online content, consumers share their experiences and assist prospective consumers in their decision-making. Additionally, understanding these insights allows organizations to develop new strategies and address new challenges for various areas of the enterprise, starting with marketing as well as quality management directed toward finance along with human resource management (Aral, Dellarocas, and Godes, 2013). The power of social media extends beyond marketing and also aspects about consumer behavior (Aral, Dellarocas, and Godes, 2013). By acquiring new insights and expertise from these unstructured data-generating sources (Benbya and Alstytne, 2010), organizations, for example, could also improve the design of their new products (Zwass, 2010). In order to discover insights of a specific quality concern of a product or service, previous studies developed robust methodologies for classifying the unstructured text that mention quality issues, e.g., performance and safety concerns (Abrahams et al., 2012; Adams et al., 2017; Goldberg and Abrahams, 2018; Law et al., 2017; Nasri et al., 2018).

1. Prior Work and Research Gaps

Prior text classification studies are generally focused on creating accurate classifiers, but the three chapters in this dissertation look specifically at the following areas: (1) identifying text that describes specific types of safety concern incidents across multiple product categories, (2) after classifying text, assessing the level of risk that the coded text implies for products, (3) after classifying text, quantifying the association of certain text codes (categories) with customer

satisfaction. In doing so, the dissertation will extend text classification literature to offer actionable insights to organizations.

1.1. Online Review Analytics

Online reviews are an unstructured data source in that text does not conform to fields (tabular columns) like traditional data sources. When a consumer posts a review online, the review may contain a blend of various modes of expression, such as emotive feedback ranging from positive to neutral to negative. These modes of expression could even describe non-emotive feedback on product usage by the consumer. Past research revealed that it is difficult to spot specific product-related issues, as the consumer-generated text is quite dispersed in 3rd party e-commerce websites such as Amazon, eBay, and so forth. Additionally, the text may have a factual tone for sharing domain-related problems such as performance and safety concerns. As the wording of the defect-specific reviews may be less polarizing than that in other online reviews, these reviews may contain more non-emotive words than emotive words. Hence, past research has shown that conventional sentiment analytic methods are less effective in detecting product defect and safety concerns (Law et al., 2017; Nasri et al., 2018).

Sentiment classification approach as well as opinion mining originate from the equivalent methodologies such as natural language processing (NLP), information retrieval (IR), information extraction (IE), as well as artificial intelligence (AI) (Xiang et al., 2015). According to Schmunk et al. (2013), “typical tasks of sentiment analysis include (1) finding documents relevant for a specific topic or purpose; (2) pre-processing collected documents, e.g., tokenizing documents into single words and extracting relevant information from them; and (3) identifying the sentiment surrounding the product or company.” Sentiment analysis is

one type of text mining that is useful in identifying subjective statements that contain opinions and emotion taken from UGC. However, there is still considerable difficulty in retrieving new knowledge using these common techniques when online review content contains category-specific information.

1.2. Text Mining and Text Classification

In classifying text, Abrahams et al. (2012) used a traditional sentiment dictionary, General Inquirer (GI) to analyze consumer comments from online discussion forums pertaining to vehicles. The researchers concluded that sentiment analysis was insufficient for finding and categorizing vehicle defects reported by consumers. Instead, they created a set of linguistic markers, known as “smoke terms”, found in discussions such as vehicle consumer forums. The “smoke terms” were more predictive of automotive safety concerns or performance defects than sentiment words.

While the consumer-generated content has focused on studying consumer complaints, there have been very few studies (Adams et al., 2017; Law et al, 2017; Nasri et, al. 2018) examining product quality issues such as defect existence and safety hazards across multiple product categories being purchased by consumers through online media or from a third-party website (e.g., Amazon.com). The current studies in *Chapter 1* and *Chapter 2* adapt the semi-automatic text classification techniques related to the previous work (Abrahams et al., 2012; 2013; 2015) in safety identification across multiple domains, extracted from online reviews. The findings can benefit the product designers as well as regulators to identify the most prevalent category-specific safety concern terms across multiple product categories. Ultimately, our risk-assessment-scoring model will help support decision-making for

stakeholders such as compliance officers, risk-management officers, and other risk-prevention experts; the framework will help them decide whether they should mitigate the risks or hazards related to a specific product.

Chapter 3 extends multiple past studies (Hawkins et al., 2016; Litvin et al, 2008; Xiang and Gretzel, 2010; Xiang et al., 2017) where service quality issues are investigated through online reviews, focusing on a particular industry (e.g. hospital, hotel, restaurant, etc.). The study in Chapter 3 outlines a method for determining various service code types mentioned in online reviews and understanding their association with online star ratings of hospitals. The findings can help service operation practitioners and academics to identify and address specific types of customer concerns that are most strongly associated with overall satisfaction.

In this dissertation, two major types of quality concerns are discussed: product safety concerns and treatment service concerns. In the following sections, the sub-categories of these concerns are defined and discussed. An overall research methodology is presented, and then the three chapters that comprise this proposal are discussed in detail.

2. Definitions

To clarify the terminology we utilize, appropriate definitions for each core term are provided below, along with linkages to the relevant chapters:

2.1. Safety concern subcategories (Chapter 1)

This study focuses on consumer reviews that mention issues causing or foreseeably causing serious injury or death. The categories of the safety concerns are broken down into event,

hazard, and injury subcategories. Each subcategory consists of multiple *code types*. The *event category* code type consists of code values such as potential injury, treatment, and actual injury. The *hazard category* code type consists of code values such as choking, falling, fire, and so forth. For instance, a reviewer mentioning that a product catches fire would be a safety issue as a fire hazard could foreseeably cause serious injury or death. The *injury category* code type consists of code values such as asphyxia, bruise, burns, and cut.

2.2. Risk Assessment Model (Chapter 2)

This study motivates the researchers to incorporate the hazard-specific reports extracted from online (real-time) reviews in a *risk assessment model* for revealing the specific safety concerns of products. Along with the *risk metrics* adapted from the practitioner's risk model, our model uses the semi-automatic text classification procedure in identifying and assessing the *risk level* of different hazards about consumer products. Theoretically, researchers indicate that risk assessment is divided into two different methods: “*formal*” as well as “*informal*” risk assessment methods. Apart from using the informal risk assessment, the formal risk assessment is the most popular methodology used among the practitioners such as government agencies because this method uses quantitative risk metrics such as the *likelihood of injury* incidents and *severity level of injury*. This paper combines these existing risk metrics with the text classification method (from Chapter 1) to construct our risk assessment model that categorized the risk level of product safety present in online reviews.

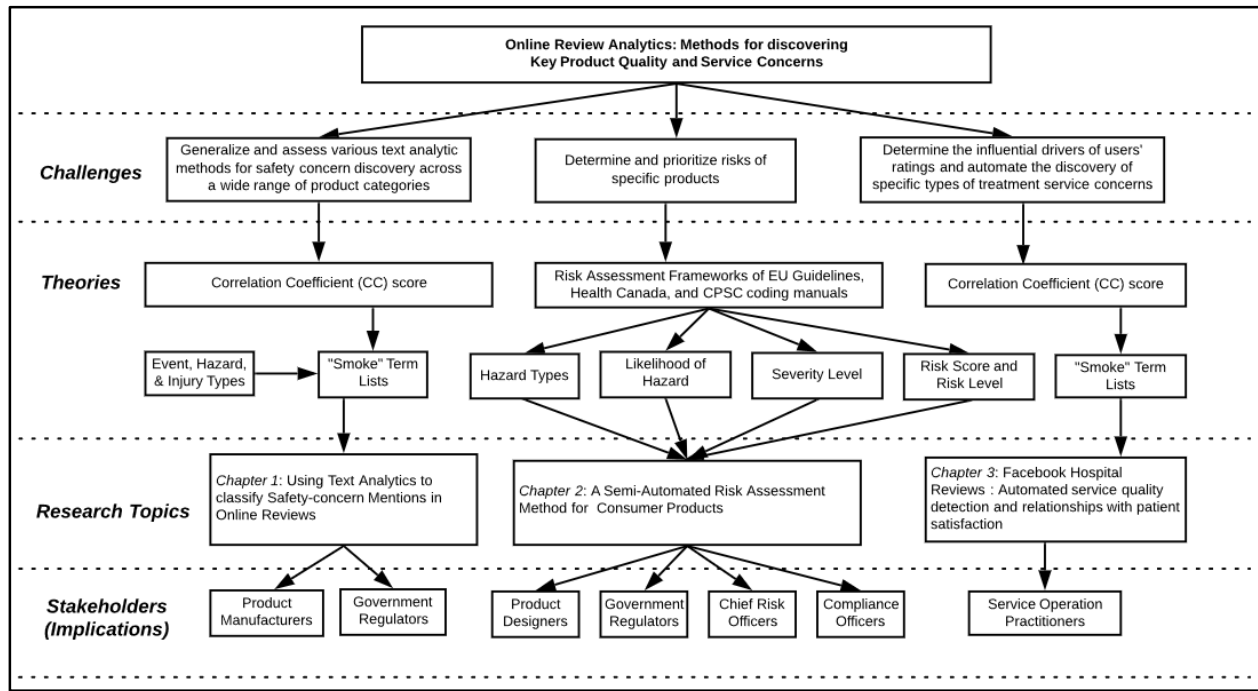
2.3. Hospital Service Codes (Chapter 3)

This study aims at classifying patient reviews into different categories of hospital *service code types* such as communication, waiting time, environment, incident type, treatment, diagnosis, and cost that may determine online star ratings. Each code type contains multiple *service code values*. For instance, a treatment code type consists of two different service code values such as harmful treatment and helpful treatment. Each of these categories and their corresponding service code values capture patients' perceptions and experiences that are associated with online star ratings of hospitals. This study found that a subset of the service code types were significant antecedents of online user ratings.

3. Text-Analytics-Based Methods for discovering Key Product Quality and Service Concerns in Online Reviews

Product quality issues such as safety concerns as well as hospital service issues may be described in a factual tone. This dissertation proposes a text-analytics-based method (see Figure 1) to measure online content and extract prevalent terms (i.e. “smoke terms”) for product safety concerns and treatment service quality concerns. The method incorporates reviews that can be used to discover and analyze concerns as well as various subcategories of serious safety concerns across multiple products. Additionally, the method analyzes patient reviews from hospital Facebook pages to identify different categories of hospital service codes, generate “smoke word” lists of each code value, and predict overall star ratings. This proposed method can present in the act of a model for quality management (QM) professionals, regulators, product designers, and service operation practitioners for evaluating unstructured user-generated content in online reviews.

Figure 1. Text-Analytics-Based Methods for discovering key Product Quality and Service Concerns



Each chapter of this dissertation addresses a specific research gap, methodology, and practical implications for specific stakeholder types. This dissertation involves semi-automated text classification procedures, as well as concentrating on product quality concerns across diverse product categories and service concerns. The contribution of each chapter is developing as well as validating (testing) the three methodologies.

Chapter 1 compares the text-classification methodology with the traditional sentiment analysis in classifying safety concerns of hazardous products into various categories.

Furthermore, past text classification studies look only at binary detection (existence vs. not) of safety concerns, whereas the current study uses text classification to disentangle multiple different subcategories of safety concern. Hence, companies can rapidly find specific types of safety concerns from real-time data through online reviews. Past studies mentioned that

practitioners such as product safety experts and product designers are unable to monitor every product on the market and often takes action on the products with safety concerns after the events have already taken place. Contrary to past studies, this chapter aims to generate “smoke” terms in the rapid discovery of specific subcategories of safety concerns from the real-time reviews.

Chapter 2 presents and evaluates a method that extends practitioners’ risk assessment frameworks. These safety hazard risks are quantified as the likelihood of causing a specific hazard code type, with particular hazard severity. This chapter proposes a semi-automated risk assessment framework to help businesses characterize the risk level of products specific to a particular hazard. Determining these risk levels of various hazard reports of using consumer products will help benefit businesses in making decisions related to the risk management of their products.

Lastly, *Chapter 3* presents and tests a method for discovering the association between major treatment service concern types and overall ratings of hospitals. The findings of this chapter help organizations to identify and prioritize key treatment service code types that impact overall user ratings. This chapter evaluates the key treatment service code types to find treatment service concern types that have the most impact on overall customer satisfaction. Additionally, this chapter extends previous studies by generating “smoke terms” for a subset of treatment service code values that are useful for health care providers to rapidly discover specific treatment service concerns in order to improve their hospital services.

References

- Abrahams, A. S., W., Fan, G.A., Wang, Z., Zhang, J., Jiao. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management* 24(6), 975–990. (doi: 10.1111/poms.12303).
- Abrahams, A. S., J., Jiao, W., Fan, G.A., Wang, Z., Zhang. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems* 55(4), 871-882.
- Abrahams, A. S., J., Jiao, G.A., Wang, W., Fan. (2012). Vehicle defect discovery from social media, *Decision Support Systems* 54(1), 87-97. (doi: 10.1016/j.dss.2012.04.005).
- Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International journal of medical informatics*, 100, 108-120.
- Aral, S., Dellarocas, C., & Godes, D. (2013). Introduction to the special issue—social media and business transformation: a framework for research. *Information Systems Research*, 24(1), 3-13.
- Benbya, H., & Van Alstyne, M. (2010). How to find answers within your company. *MIT Sloan Management Rev.*, 52(2), 65-75.
- Goldberg, D. M., & Abrahams, A. S. (2018). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*, 105, 52-65.
- Hawkins, J.B., J.S., Brownstein, G., Tuli, T., Runels, K., Broecker, E.O., Nsoesie, D.J., McIver, R., Rozenblum, A., Wright, F.T., Bourgeois, F., Greaves. (2016). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety* (October), 1–10 (doi: 10.1136/bmjqs-2015-004309).
- Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67, 84-94.
- Litvin, S.W., R.E., Goldsmith, B., Pan. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management* 29(3), 458-468.
- Nasri, L., Baghersad, M., Gruss, R., Marucchi, N. S. W., Abrahams, A. S., & Ehsani, J. P. (2018). An investigation into online videos as a source of safety hazard reports. *Journal of Safety Research*, 65, 89-99.
- Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2013). Sentiment analysis: Extracting decision-relevant knowledge from UGC. In *Information and Communication Technologies in Tourism 2014* (pp. 253-265). Springer, Cham.

- Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58, 51-65.
- Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management*, 31(2), 179-188.
- Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management*, 44, 120-130.
- Zwass, V. (2010). Co-creation: Toward a taxonomy and an integrated research perspective. *International journal of electronic commerce*, 15(1), 11-48.

Chapter 1. Using text analytics to sub-classify safety-concern mentions in online reviews

Abstract

In this study, we extend past studies examining safety surveillance of online reviews by adapting standard injury severity coding schemes from the medical field. We examine online reviews reporting specific categories of safety concerns and contrast them with reviews not reporting these safety concerns. We utilized a large dataset of Amazon.com reviews covering diverse product categories. Each review was manually labeled (coded or “tagged”) to indicate whether it referred to a safety concern, and over 700 reviews out of 124,289 examined were determined to refer to safety concerns. We developed event-specific, injury-specific, and hazard-specific “smoke” term lists, or sets of unique words and phrases indicative of these linguistic patterns. These smoke terms were used for detecting, categorizing, and prioritizing safety concerns across product categories. By utilizing unsupervised topic modeling and supervised smoke term techniques, we compared lists for overlap to determine which terms were generic and which were specific to safety concerns within certain domains. Additionally, we tested the generalizability of using the smoke term scoring methods in discovering the category-specific safety reports by testing our findings on a new dataset collected from a Fortune 50 retailer’s online reviews. This study has benefits for organizations and safety regulatory agencies to prioritize and triage safety concerns for more detailed analysis of online reviews.

1. Introduction

Recalls usually take place due to a failure in quality standard or due to health and safety risk, which the manufacturer did not specify before selling the product to the consumers. A safety recall of a product may occur when a manufacturer of the product discovers a safety concern in the product while selling it to the consumers (Shaaban, Maher, and Singhapakdi, 2019). Hora, Bapuji, and Roth (2011) mentioned industries including automobiles, pharmaceuticals, medical devices, electronics, toys as well as food usually declare recalls to products which could result in safety issues. According to Hora, Bapuji, and Roth's (2011) estimation, average time to recall a defective product ranged from 436 days to 869 days. Yao and Parlar (2019) integrated real-time customer feedback on products for developing a recall timing decision process. They predicted the length of the time that a product stays on the market prior to its recall date. Additionally, they also used the recall timing decision process to improve the prediction of "product defect rate", so that they could establish an improved approach in estimating the costs (Yao and Parlar, 2019).

The Consumer Product Safety Commission (CPSC)¹ found that occurrences related to safety concerns of products cost the country above \$1 trillion every year². Nasri et al. (2018) mentioned that CPSC is unable to monitor every product on the market and often acts on products with safety concerns after hazardous events have already taken place.

This paper focuses on classifying safety concerns of hazardous products into various categories so that companies and regulatory agencies can rapidly find specific types of safety concerns from real-time data through online reviews to avoid or reduce the time in product recalls.

¹CPSC. (2019). About CPSC. Retrieved from <https://www.cpsc.gov/About-CPSC>

² <https://utahfamily.com/main/>

Online reviews have recently become a common platform for firms and regulators searching for safety concern reports. As a large population of consumers report their daily product-related feedback and experiences in the form of online reviews, stakeholders such as regulators may unravel these reports for monitoring safety concerns. However, it is cumbersome to identify safety concern reports as there only a small percentage of online reviews mention these safety concerns. Manually reading through millions of reviews to locate only a tiny portion of safety concern reports for each product category is very difficult and time-consuming for stakeholders.

Our goal is to apply text analytics on online reviews by discovering specific words and phrases (terms) with remarkably high frequency in safety concerns relative to non-concerns. Prior text classification research typically looks only at safety concern existence vs. non-existence, in a binary fashion, and does not investigate safety concern subtypes. We go beyond prior research, by identifying terms peculiar to specific safety concern narrative subcategories: event, hazard and injury types. Terms with unusually high prevalence in safety concerns are called “smoke” terms in past studies (Abrahams et al., 2012). In the smoke word-based approach, researchers detect n-grams (1, 2, and 3-word phrases) that are significantly more common in the target category of interest (e.g., safety concern). These n-grams are found using supervised learning on human-tagged data. Each smoke term (word or phrase) is assigned a weight based on a prevalence metric (Fan et al., 2005), and then new (unseen) texts are scored based on the sum of the “smoke” terms present in the n-gram dictionaries. This study adapts this method that is proven by a series of previous studies in detecting overall defects (i.e., both performance and safety concerns) as well as specifically safety concerns in social media among various consumer products (Abrahams et al., 2015; Abrahams et al., 2012; Adams et al., 2017;

Goldberg and Abrahams, 2018, Law et al., 2017; Winkler et al., 2016). Prior research (Abrahams et al., 2015; Adams et al., 2017; Mummalaneni et al., 2018; Nasri et al., 2018; Winkler et al., 2016) related to safety concern classification is typically binary and identifies only the existence of a safety concern. We apply this human-machine system to a dataset of reviews from the United States' largest online retailer, Amazon, by examining category-specific terms that indicate prevalence of event types, hazard types, and injury types across diverse product categories. Additionally, this study aims to show the generalizability of the “smoke” terms by discovering the category-specific safety concern reports not only from Amazon.com reviews but also from a different data source derived from a Fortune 50 retailer based in the U.S. By utilizing the methodology of this current research, firms and regulators such as CPSC can automatically prioritize and routinely respond to the specific types of safety concern narratives. This study also aims to evaluate the performance of human-machine classification method with that of traditional sentiment analysis, demonstrating the superior performance of machine-tuned smoke terms.

2. Literature review

This section reviews previous work in safety concern discovery that uses conventional structured data sources, online reviews, and text analytics.

2.1. Structured data sources

Government agencies such as the CPSC, Health Canada, European Union (EU) Health and Safety Authority, and British Standard Institution (BSI) identify and evaluate risks at different stages of the product safety cycle. These federal agencies keep archived narratives of various

safety concerns related to multiple consumer products. They regulate the ones that pose a severe safety concern to consumers.

Saferproducts.gov has become an important database for consumers to report hazard incidents. The website was authorized by Congress in 2008, went active in 2011. Over the past few years it has become progressively accessible to the public. In 2014, the CPSC mentioned that they received approximately 20,000 reports of product-related harm or risk (CPSC, 2014). Congress is currently considering improvements to saferproducts.gov. Congressional testimonies have been presented by the Consumer Federation of America³, Consumers Union, Kids in Danger (KID), Public Citizen, the U.S. Public Interest Research Group (U.S. PIRG) and others. These stakeholders have recommended that the CPSC to merge additional data sources and resources into saferproducts.gov and increase the data analysis for enhancing the categories of harm and hazards that are listed in the database (Gregg, 2019).

This study uses unstructured data to detect and categorize reports of safety concerns that CPSC should know about and may help identify product categories that the agency and manufacturers should be aware of that are most likely to cause hazards to the consumer's safety. Even though the unstructured data is difficult to analyze in depth due to its volume, the stakeholders and regulators may benefit from monitoring its real-time information.

³ <https://consumerfed.org/>

2.2. Online reviews

Online reviews contain a mix of positive, neutral, and negative comments on various products. As there are so many online reviews, it is challenging to separate a distinct problem of a product category from other textual content. Prior studies investigated fragmented online reviews to identify linguistic trends. The wording of the defect reports may be less emotively polarizing than that in other online reviews (Mummalaneni et al., 2018). Likewise, the text of the online reviews explaining safety concerns is frequently factual, rather than polarizing. Hence, it becomes difficult for traditional text analysis to locate the safety issues. For instance, online consumers frequently share their emotions such as negative or positive sentiments but seldom discuss their safety concerns of a specific product that could be susceptible to a specific type of injury.

2.3. Sentiment analysis and text classification

Sentiment analysis applies methods from natural language processing to automatically detect the polarity of author's attitude toward a topic: positive, negative, or neutral (Wang & Wang, 2014). Sentiment analysis can be "coarse-grained" or "fine-grained". Whereas "course-grained" sentiment analysis alone detects polarity (positive or negative), "fine-grained" sentiment analysis detects both the polarity and strength (Jaimes et al., 2006; Kanayama and Nasukawa, 2012). Sentiment strength may be dissimilar for words with the same polarity. For example, "good" and "excellent" are both classified as positive, but the latter states stronger sentiment than the former (Wang and Wang, 2014). Many approaches are dictionary-based. For example, positive as well as negative terms can be kept in a word dictionary (Nielsen,

2011; Abrahams et al., 2015) such that each word is associated with sentiment strength within some range, such as +5 to -5. For example, the OpinionFinder system uses lexicons to detect polarity direction for a subject or a target class (Abrahams et al., 2015).

Even though multiple research studies used text classification techniques to enhance decision-making strategy for businesses, these methods were useful to identify product quality issues (Pan et al., 2014; Vallmuur et al., 2016). Extensive investigation is needed to utilize the prior studies for increasing the performance of the text classification techniques. Previous studies used text classification approaches to discover defect in the automotive industry (Abrahams et al., 2012) and then the subsequent study developed a framework for detecting defect by using text classification approaches across multiple industries (Abrahams et al., 2015). Later studies applied this defect-discovery-framework to untangle safety concern incidents of toy products (Winkler et al., 2016), performance defects of dishwasher appliances (Law et al., 2017), and safety concern reports of baby cribs from the online reviews (Mummalaneni et al., 2018). In this study, we aim to adapt from the previous studies, with a significant focus on sub-classifying the safety concern categories among various product categories.

2.4. Research contribution and research gap

Table 1 summarizes previous academic studies related to safety surveillance. These studies either used unsupervised learning approaches or combined machine-human intelligence for discovering safety concerns. For example, Marucci-Wellman et al. (2017) compared performance of a combination of human and machine approaches to classify short injury

narratives, containing only up to 120 characters. Previous studies narrowed down to specific product categories, except for Nasri et al. (2018), who investigated online videos across product categories to detect the existence of safety concerns. Vallmuur et al. (2016) mentioned future research towards building and testing methods for semiautomatic classification by using human-machine learning of injury reports, applicable to more than one injury domain. Instead of classifying only short injury narratives (Marucci-Wellman et al., 2017), this study aims to classify ‘big injury narrative data’ (Vallmuur et al., 2016) into sub-categories of safety concerns among multiple product categories by using a machine learned classification approach. This sub-classification technique may allow us to generate useful results for multiple firms and regulators, instead of only focusing on a specific product category.

Study	Product category	Data source	Sample details	Binary safety concern vs Sub-classify safety concern types
Abrahams et al. (2012)	Automotive	Automotive enthusiast discussion forums	4,500 discussion threads	Binary
Adams et al. (2017)	Joint and muscle pain relief medicine	Amazon.com	32,518 online reviews	Binary
Alemzadeh et al. (2013)	Medical devices	FDA	1,210 recall reports	Binary
Boutros et al. (2014)	Food	Questionnaire, interviews	5 questionnaires, 3 interviews	N/A
Marucci-Wellman et al. (2017)	Worker	Workers' compensation claims	30,000 cases	Sub-classify
Mummalaneni et al. (2018)	Baby cribs	Amazon.com	12,742 online reviews	Binary
Nanda et al. (2016)	Workplace products	Survey of Occupational Injury and Illness	50,000 injury cases	Sub-classify
Nasri et al. (2018)	Cross-category	YouTube.com	15,402 video reviews	Binary
Pan et al. (2014)	Electric products	CPSC	3,192 safety reports	N/A
Shi et al. (2017)	Aviation operations	Aviation Safety Reporting System	168,227 aviation incident reports	Binary
Winkler et al. (2016)	Toy	CPSC NEISS, CPSC recalls	587 injury reports, 1,065 recall reports, 1.05 million online toy reviews	Binary
Our study	Cross-category	Amazon.com	124,289 online multiple product reviews	Sub-classify
		Fortune 50 Retailer (2 nd Data Source)	36,488 reviews of seasonal products	

Table 1. Survey of academic safety hazard surveillance literature.

Past studies (Abrahams et al., 2015; Abrahams et al., 2012; Adams et al., 2017; Goldberg and Abrahams, 2018; Law et al., 2017; Mummalaneni et al., 2018; Nasri et al., 2018; Winkler et al., 2016) undertake safety concern classification in a simple binary form ("safety concern exists" or "safety concern does not exist"). Zhang et al. (2019) employed an unsupervised defect classification methodology for detecting product performance defects. Thus, they have not specifically investigated safety concerns as they identified domain-specific automotive product issues such as component, symptom, and resolution, which are non-binary defect classifications. This study extends the prior works by classifying various sub-categories of safety concerns by using an unsupervised learning technique, topic modeling, as well as by using supervised learning approaches. To enhance the domain-related nature of our text mining outcomes, we used the human-machine learning classification approach in our current study. Natural Language Processing (NLP) techniques are utilized to

further classify these safety concern reports into more specific sub-categories than were studied in these past papers (Abrahams et al., 2012; Abrahams et al., 2015). For example, we added some additional codes (treatment, injury mitigation, product remediation) that do not appear in the past papers, and developed machine learning (ML) classifiers, that do not appear in the past papers, for all codes (sub-categories). Before the computer-aided algorithm generated the distinct keywords for each sub-category, we performed a human coding (i.e., tagging) procedure that we will elaborate next, in the Methodology section.

3. Methodology

The following subsections describe the data source, data coding, and data processing.

3.1. Data source and data coding

For this study, we utilize online reviews collected from McAuley et al.'s (2015) Amazon review data. This dataset⁴ contains the online reviews that span May 1996 - July 2014. We assembled the dataset consisting of reviews from 17 product categories as shown in Table 2. Table 2 shows 17 product categories that we used in our study. These 17 product categories represent a broad cross-section of retail products available for purchase online (see Table 2). Thus, they allow generalizability as to product safety concern occurrences across industries. Table 2 demonstrates whether the datasets for each product category were randomly sampled from McAuley's dataset², or by first scoring McAuley's data² using correlation co-efficient (CC) score (Fan et al., 2015) or sentiment techniques like AFINN (Kuiper and Hood, 2019).

⁴ McAuley, J. Amazon product data. Retrieved from <http://jmcauley.ucsd.edu/data/amazon/>

Product category	Total tags	Unique tags	Reviews with Safety concerns
Baby Products	27981	16930	850
Random	22331	13669	627
Scored	5650	3261	223
Blenders	21095	16869	577
Random	19894	15768	501
Scored	1201	1101	76
Car Seats	22635	20000	1438
Random	22635	20000	1438
Clothing	6782	3557	111
Random	3902	2055	34
Scored	2880	1502	77
Dishwashers	6052	4043	56
Random	3216	2309	37
Scored	2836	1734	19
Elderly Products	15259	14443	617
Random	15259	14443	617
Furniture	7643	6674	48
Random	6030	5261	32
Scored	1613	1413	16
Garden Tools	6245	4065	70
Random	3815	2066	44
Scored	2430	1999	26
Household Products	7705	3612	104
Random	4240	2140	46
Scored	3465	1472	58
Musical Instruments	1605	1405	3
Scored	1605	1405	3
Office Products	6503	3185	18
Random	3930	2057	14
Scored	2573	1128	4
Power Tools	6176	3050	189
Random	3149	1996	49
Scored	3027	1054	140
Refrigerators	6173	4742	32
Random	3510	2918	23
Scored	2663	1824	9
Small Appliances	11450	5519	97
Random	6825	3755	65
Scored	4625	1764	32
Smartphones	5700	4254	24
Random	3185	2483	15
Scored	2515	1771	9
Sports Equipment	7710	4807	229
Random	4860	3210	133
Scored	2850	1597	96
Toys	15285	7134	475
Random	12885	5533	443
Scored	2400	1601	32
Grand Total	181999	124289	4938

Table 2. Amazon Dataset: Total reviews and unverified safety concerns in 17 product categories

The data sampling and scoring were undertaken by undergraduates. While the scoring may be imperfect, we nevertheless consider these datasets to be a fertile set of unverified safety hazards. In total, 124,289 unique reviews (Table 2, Column 3 grand total) were analyzed across these 17 product categories. Teams of undergraduate business students at a large public research university were tasked with tagging (“coding”) these online reviews, or reading each review and labeling whether or not it referred to a product safety concern. We asked students to tag about 200 reviews each. These reviews were assigned to each student randomly; in total, the students generated 181,999 tags (Table 2, Column 2 grand total; some reviews were tagged more than once). If any of the taggers tagged a review as referring to a safety concern, we gave the review an initial designation of “safety concern” until further review (Goldberg and Abrahams, 2018). 4,938 reviews, or 3.97% of the total, were tagged as referring to a safety concern (Table 2, Column 4 grand total).

To verify the tagging accuracy of the students, 4,938 reviews were randomly distributed to a second group of undergraduates, who indicated whether they believed a safety concern was present. Of the 4,938 reviews, the second round of undergraduates confirmed only 1,389 safety concern reports. Then in a third round, graduate students (including the lead author) carefully reviewed the 1,389 reports by removing duplicates and incorrectly tagged-reports. At the end, they verified that 740 out of the 1,389 reports truly referred to safety concern narratives. Table 3 (Amazon reviews sub-total) shows the breakdown by product category for the 740 safety-concern reports that were utilized in the analysis detailed in the remainder of this chapter.

For the purposes of validating our method, we also obtained a second dataset, comprising 36,488 reviews for seasonal products (see Table 3), from another large online retailer in the United States. We refer to the 2nd large online retailer as “Fortune 50 retailer”, due to confidentiality restrictions.

The Fortune 50 retailer dataset contains online reviews of 5,569 unique products, across 502 product categories, all subcategories of “Seasonal items”. From the dataset of Fortune 50 retailer, there are a total of 42,990 tags (36,488 unique reviews). The agreement percentage between the graduate student authority taggers and undergraduates was 86.3%, which comes out to a Cohen's K of 0.73, representing “substantial” inter-rater agreement, according to (Landis and Koch, 1977). The number of overlapping cases was 1,398 where there were 1,207 agreements (i.e. 86.3%). For verifying the tags, two graduate students (including the lead author) tagged 1,079 reviews for comparing their tags to the undergraduate students’ responses. The agreement percentage between the two graduates was 93.8% (Cohen’s K = 0.88) where there were 65 overlapping cases. We found that 666 out of 36,488 Fortune 50 retailer reviews (i.e. 1.8%) were tagged as safety concerns during the verification by the graduate students.

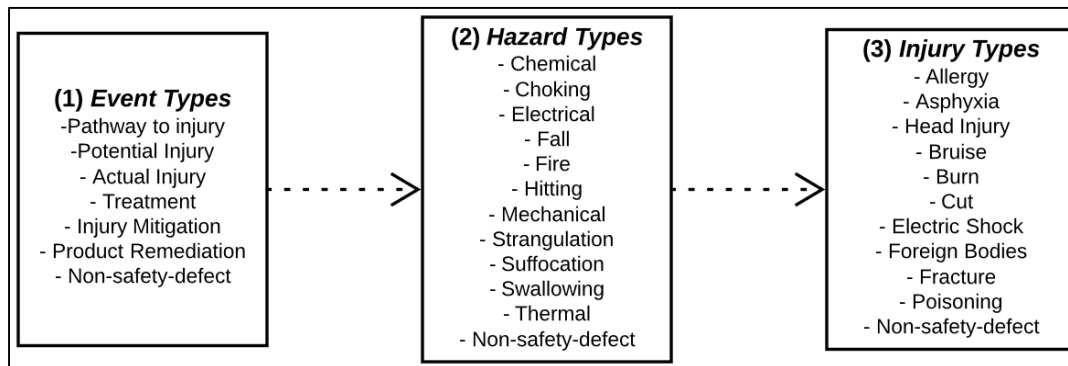
Data Source	Product categories	Reviews with verified safety concerns
Amazon	Appliances	18
	Baby	380
	Clothing, Shoes, and Jewelry	9
	Health & Personal Care	142
	Home & Kitchen	54
	Patio, Lawn, & Garden	10
	Sports & Outdoors	14
	Tools & Home Improvement	7
	Toys & Games	106
	Sub-Total	740
Fortune 50 retailer	Seasonal items	666

Table 3. Verified safety concerns by product category.

As the literature acknowledges (Mummalaneni et al., 2018), the proportion of online discussions mentioning safety concerns explicitly is quite small. As such, this enormous data collection effort is necessary to generate a sample size sufficient to study this cross-category problem,

This study aims to further classify the safety concern narratives in order for this human-machine classification method to filter safety reports in real-time. There are multiple different ways to look at safety concern categories. For instance, the Health Canada coding manual looked at risk posed by event type, different hazard types, and different kinds of injury narratives (Canada, 2016), whereas CPSC assesses risk in terms of hazard identification (Poston, 2018). Hence, this study has adapted these multiple ways to code safety concern subcategories: event, hazard, and injury types from the coding manuals of CPSC and Health Canada. Figure 1 illustrates the code values within each of the following three code categories: event type, hazard type, and injury type.

Figure 1. Code values within each safety concern subcategory code type



The safety concern reports were manually classified (tagged) for each code type.

Within these safety concern reports, the human taggers tagged different code types within the safety-related categories. Figure 1 shows that safety-related (1) event types are described in online reviews. These reviews may describe specific (2) hazard types. Then these hazard types sometimes may lead to particular (3) injury types. In the event type, examples of the code values are pathway and actual injury. In the hazard type, examples of the code values are choking and fire. In the injury type, examples of code values are head injury and bruise injury. Each of the event, hazard, and injury types consist of multiple code values. For instance, the categorical (tag) type “Hazard Type” may consist of the code values “Choking”, “Fire”, “Strangulation”, and so forth, or the categorical type “Event Type” may consist of the code values “Pathway” or “Actual Injury”. Categorical types relate to column heading names, while code values relate to cell values in tabular format.

For classifying the code types within each safety concern category, we focused on manually tagging every sentence of the 740 Amazon.com reviews that were verified as being referred to safety concerns, which comprised 5,499 total sentences. Additionally, we manually tagged 1,958 total sentences (i.e. 666 safety-concern reviews) from the data set of

the Fortune 50 retailer. Each sentence of the safety-concern reviews collected from the two data sets (Amazon and Fortune 50 retailer) was tagged separately as to its event type, hazard type, and injury type.

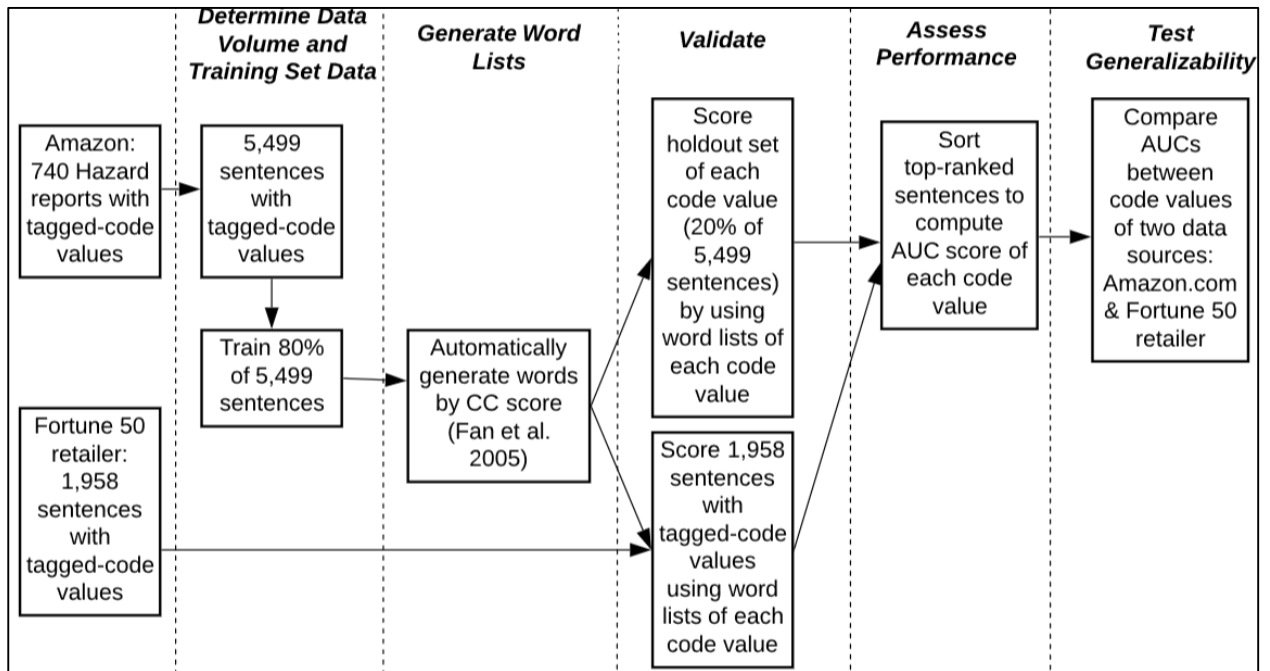
For each tag, the tagger was presented with the full text of the review so that the tagger could view the full context, and one focal sentence in question was emphasized. We devised a tagging protocol (see APPENDIX A) in which each possible event type, hazard type, and injury type is described. Two taggers were tasked with analyzing this dataset further, which we refer to as tagger A and an authority tagger B. Each tagger individually tagged a random set of sentences for each code value within each safety concern category. The tagging protocol contains a detailed explanation describing each code value for sharing with the taggers A and B. The authority tagger B tagged all the 5,499 sentences to classify these code types. Then the tagger A tagged several random sets (see Table 3) to classify these code values, as necessary, for validation purposes to check whether the protocol was clear enough to follow the tagging instructions. To determine inter-rater agreement (i.e., reliability) between the taggers A and B, we computed Cohen's Kappa of each code type (Landis and Koch, 1977). After the tagger A tagged the first set of 200 sentences from the Amazon.com dataset to classify each code type, the Cohen's Kappa scores were 0.63, 0.82, and 0.66 for the event type, hazard type, and injury type respectively. After tagging the last set of 200 focal sentences from the Amazon dataset, the scores were 0.75, 0.85, and 0.98 for the event type, hazard type, and injury type respectively. Hence, there was an improvement in the Kappa scores because both the taggers resolved the discrepancies in their tags. In total, about 600-800 overlapping sentences from the Amazon dataset were tagged for each category (code type) in order to gauge inter-rater reliability – the overall agreement is shown in Table 4.

After resolving the conflicts, we kept only the tags with agreement between the taggers A and B for smoke term generation. We used this same procedure on 1,958 focal sentences of online reviews from the Fortune 50 retailer based in U.S. We also observed the same trend of improvement in the tagging agreement between the taggers A and B for this second dataset, as shown in Table 4.

Source	Category	Overlapping cases	Agreement	Cohen's Kappa
Amazon	Event	805	0.76	0.72
	Hazard	805	0.85	0.84
	Injury	607	0.83	0.82
Fortune 50 retailer	Event	500	0.91	0.90
	Hazard	400	0.92	0.91
	Injury	400	0.98	0.97

Table 4. Statistics for inter-rater agreement between taggers A and B

Figure 2. Generating and Testing Accuracy of Smoke Words of each Code



Next, we performed the significant word analysis for identifying what phrases (n-grams or smoke words) are present in the individual code types. As shown in Figure 2, we used an automated approach (machine learning algorithm), which used the prevalence measure or “Correlation Coefficient” (CC) score, built by Fan, Gordon, and Pathak (2005). First, the reviews tagged with each code value were divided 80 percent into a training set and 20 percent into a holdout set from the 5,499 focal sentences of Amazon dataset. From the training set, this automated approach rapidly identified words that were especially prevalent in a specific code value. Words with high prevalence in specific code value were stored in an “n-gram dictionary” along with the CC scores that were assigned to each of the prevalent words. Each dictionary contains a set of supervised smoke terms: unigrams (single words), bigrams (2-word phrases), and trigrams (3-word phrases). After using the CC score (Fan, Gordon, & Pathak, 2005) to generate initial candidate smoke terms, we manually curated the terms to

improve generalizability: consistent with prior studies (Abrahams et al., 2012; Winkler et al., 2016), we removed common English words (e.g., and, but, it), brand-specific words (e.g. LG, Samsung, Crayola). Additionally, we removed words with CC scores less than or equal to 0.

3.2. Assessment of Smoke Words

After generating the n-gram dictionaries or word lists of various code values, we tested the performance of these scoring methods to determine the performance of these methods for delineating safety hazards in our unseen holdout set. We used each dictionary to “score” the unseen holdout set, incrementing a cumulative score for each review based upon the weights that we had identified in the previous step (Goldberg and Abrahams, 2018). After obtaining a distinct set of reports for each code value in consumer language, we tested the word lists to see whether or not we could automatically identify the safety concern reports for a specific code value within a safety concern category. Specifically, we examined whether the top-ranked reviews that were more likely to refer to a specific code value than the bottom-ranked reviews. To test the generalizability of each word list, we compared the performance of each scoring method to that of the conventional sentiment analysis on a different data set consisting of 1,958 focal sentences collected from the online product reviews of the Fortune 50 retailer. The sentiment analysis method, AFINN (Nielsen, 2011), is a lexicon-based tool that scores each sentence with values ranging from -5 (negative) to +5 (positive). A word having a negative sentiment score is assumed to indicate a higher likelihood of a specific code type existence than a word having a positive sentiment score.

4. Results and discussion

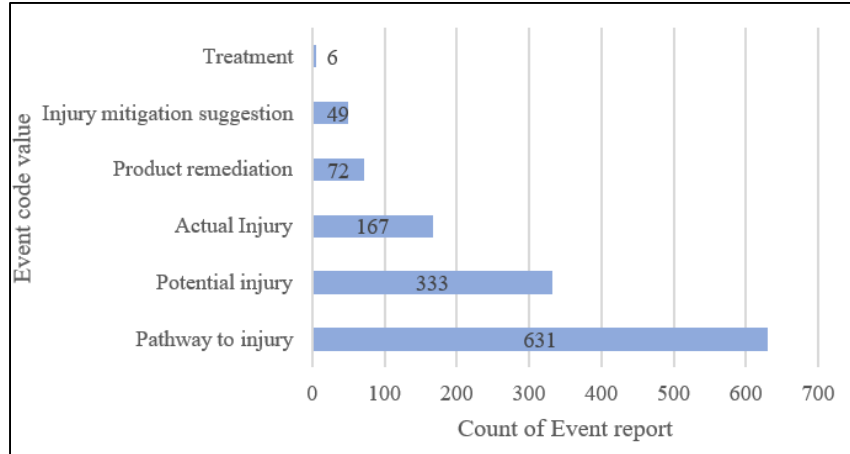
The following subsections describe the exploratory study, analysis of category-specific safety concern, and generalizability of the category-specific unique terms (i.e. smoke n-grams).

4.1. Exploratory study

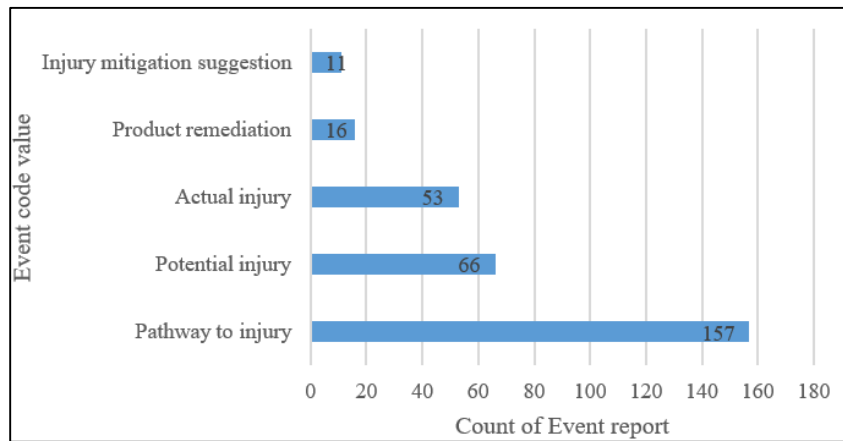
After manually tagging the full set of 5,499 focal sentences from the Amazon online reviews, we show the distribution of the existence of the event, hazard, and injury code values in charts below. Additionally, we show the distribution of all the values after manually tagging 1,958 total sentences from the online reviews of the Fortune 50 retailer. Each sentence was tagged separately as to its event type, hazard type, and injury type.

Figure 3. Count of event reports by event code value

(a) Amazon



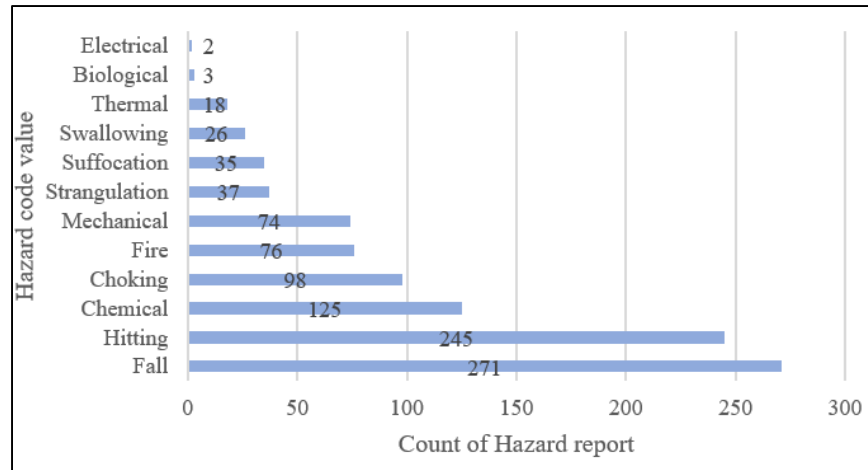
(b) Fortune 50 retailer



Figures 3-5 show that the number of reports on the Amazon dataset (a) and the dataset of the Fortune 50 retailer (b) within each safety concern category (event, hazard, and injury) correspond for each of the code value. Figure 3 shows that the pathway to injury, potential injury, and actual injury contain a large number of event reports on the Amazon dataset and the dataset of Fortune 50 retailer mentioning these specific code values.

Figure 4. Count of hazard reports by hazard code value

(a) Amazon



(b) Fortune 50 retailer

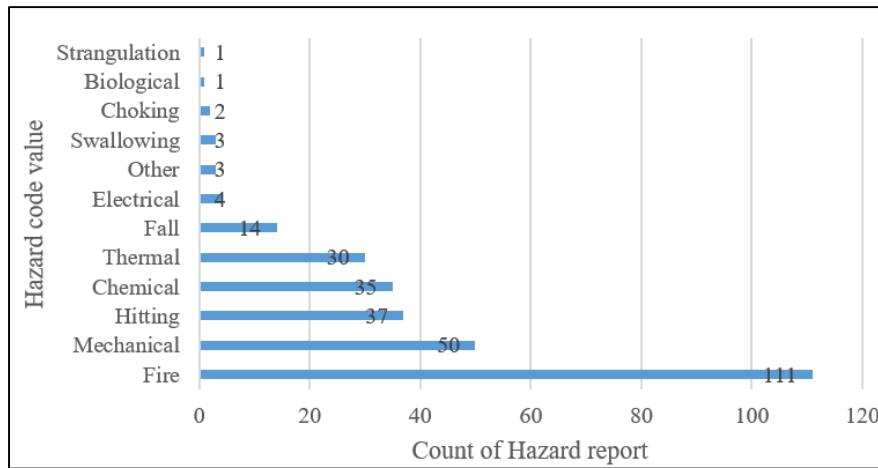
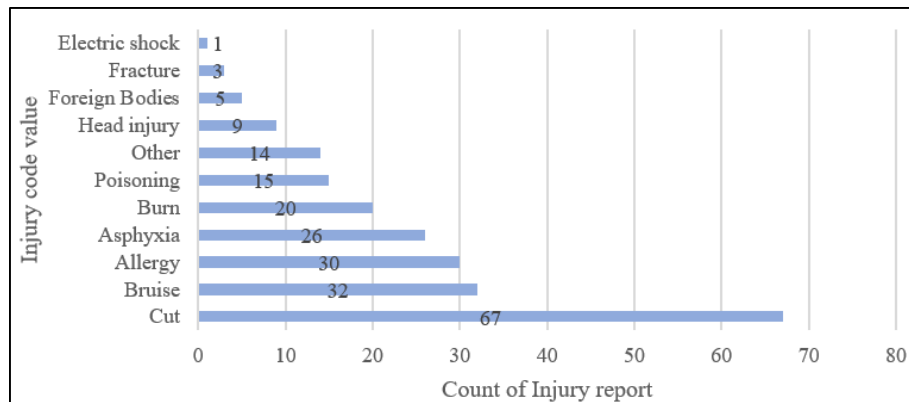


Figure 4 shows the various code values of the hazard reports. There is a high count of hazard reports that mention fall, hitting, and chemical code values in the Amazon dataset. On the other hand, there is a high count of hazard reports that mention the fire code value in the dataset of Fortune 50 retailer.

Figure 5. Count of injury reports by injury code value

(a) Amazon



(b) Fortune 50 retailer

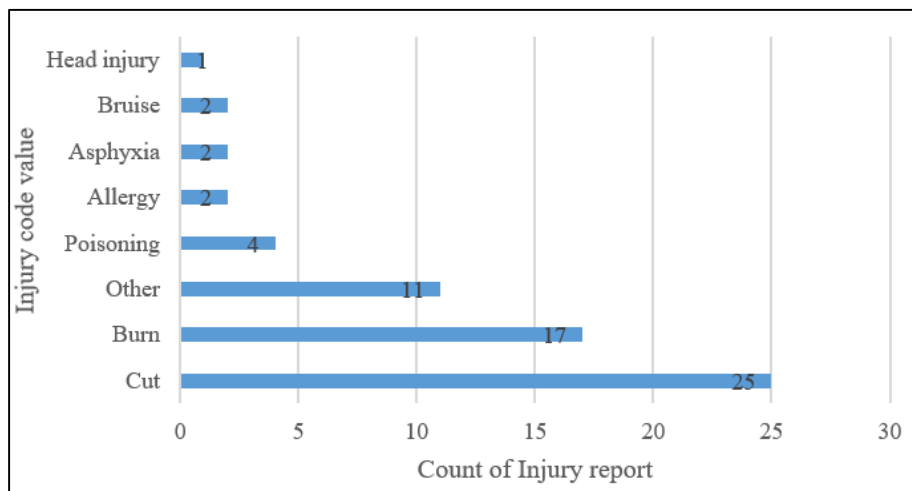


Figure 5 demonstrates Amazon's injury reports and the Fortune 50 retailer's injury reports respectively and the distribution of reports among the multiple injury code values. The cut injury code value represents the highest count of injury reports above all other code values of injury category.

4.2. Safety concern category-specific analysis

Appendix B shows the words and phrases that were most specific to each code value of the safety concern categories: event, hazard, and injury categories. The appendix contains the 30 most prevalent 1-word, 2-word, and 3-word terms (phrases) for each code value, using CC score as term prevalence metric (Fan, Gordon, and Pathak, 2005). We assessed the performance of different text analytic methods in a category-specific setting. We first included unsupervised learning technique and then began with an out-of-the box sentiment analysis technique AFINN (Kuiper and Hood, 2019) for comparing their performances with that of supervised learning approach. We used the unsupervised learning technique, LDA (Xiang et al., 2017), on 740 safety concern reports that span across a variety of different product categories. As we focused on generating terms that predicted specific code values, we ran LDA by using various numbers of topics (from five to thirty) on only the narratives that referred to safety concerns. In general, we found that as we requested more topics, LDA would return repetitive clusters that seemed to refer to the same topic. Therefore, we report in Table 4 the output obtained when using six topics and lists the top ten terms obtained for each topic. We found that LDA neatly delineated between six major types of hazards present in the online reviews: (1) choking; (2) mechanical; (3) fire; (4) fall; (5) strangulation; and (6) chemical. These code values of hazards appear to generalize fairly well across product categories, as mechanical, for example, may be present in both home and kitchen products and toys. Similarly, each product category may contain multiple code values of hazards, like toys containing both mechanical and choking hazards.

Topic	Top terms
Topic 1: choking	hazard; choking; small; pieces; little; could; plastic; lead; child; many
Topic 2: mechanical	sharp; cut; hands; edges; cut; hard; piece; finger; together; wood
Topic 3: fire	smell; fire; food; hot; time; inside; house; burning; first; never
Topic 4: fall	weight; support; fall; dangerous; loose; safety; fell; legs; floor; head
Topic 5: strangulation	strap; straps; latch; side; tighter; belt; tight; son; base; enough
Topic 6: chemical	rash; skin; dry; bad; problem; diapers; using; irritation; may; child

Table 4. Latent Dirichlet Allocation (LDA) results.

Due to the unsupervised nature of LDA, we found that the topics contained in Table 4 refer generally to only *hazard categories*. As a further analysis of category-specific safety concern reports, the supervised learning technique revealed more topics that denoted a variety of safety concern categories such as *event categories* and *injury categories* other than hazard categories. For example, some reviews may refer to events in which a potential injury occurs (e.g., “could have burned me”), while others may refer to events in which an actual injury occurs (e.g., “it cut my finger open”). Although unsupervised techniques like LDA do not lend themselves to predicting these predefined classifications of interest, we used supervised techniques such as those mentioned in Abrahams et al. (2015), Goldberg and Abrahams (2018), and Goldberg (2019) to delineate between event types, hazard types, and injury types.

Additionally, the supervised learning technique that generated the top 10 most prevalent terms (see Appendix B) in each code type had very little overlap with the sentiment words. After the machine-scoring algorithm (Fan, Gordon, and Pathak, 2015) generated the initial candidate terms for each code type, we then manually curated the terms by removing common English words (e.g., and, but, it) and brand names (Abrahams et al., 2012; Winkler et al., 2016). We also removed terms that had CC scores less than or equal to 0.

Our study discovered more non-emotive n-grams than emotive n-grams for each code value. For detailed analysis, we compared the performance of sentiment analysis with that of

the smoke term scoring methods. Appendix C summarizes the number of true positives for the specific code values for each safety concern category (i.e. event, hazard, and injury type) found in the top 200- and bottom 200-ranked focal sentences, as scored using each scoring method when analyzing the Amazon holdout set of 1,102 sentences as well as the dataset of 1,958 sentences taken from the 50 Fortune retailer. As per Table 10 (a) under Appendix C, each hazard type was scored using its corresponding hazard-type dictionary. Therefore, for example, the count in the *Top 200* row for *Unigram*, under column, *Fall* indicates the number of positives (i.e. sentences that are instances of the target class in this case, *Fall*), 42, found in the top 200 sentences after scoring with the Fall-unigram dictionary. As shown in Table 5, **bolded** numbers indicated the best-performing code value scoring methods detected a significant number of code values in the top 200-scoring sentences. The top-scoring sentences via each method were clearly differentiable from the bottom-scoring sentences, as chi-squared tests indicated significantly different proportions for each method at the 0.01 level. For instance, Table 5 shows a chi-squared test comparing the proportion of a code value, *Pathway to Injury* from the event type in the top 200-scored sentence where each scoring method was significantly different from the sentiment analysis ($p < 0.01$), as smoke n-grams detected more pathway-to-injury sentences in the top 200-ranked sentences of Amazon’s holdout set than the sentiment words.

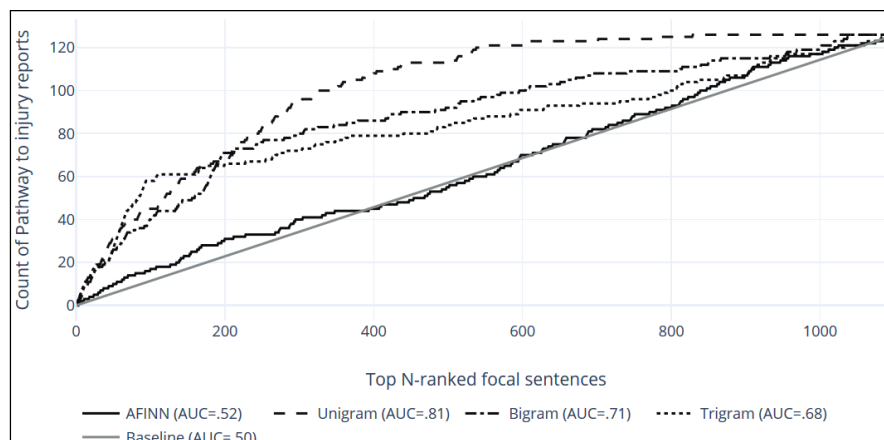
	Scoring method	Pathway to injury	No pathway to injury
Sentiment	<i>AFINN</i> Bottom 200 Top 200	21 31	179 169
Smoke list	<i>Unigrams</i> Bottom 200 Top 200	1 68	199 132
	<i>Bigrams</i> Bottom 200 Top 200	18 71	182 129
	<i>Trigrams</i> Bottom 200 Top 200	20 67	180 133

Table 5. Number of pathway to injury sentences in top 200- and bottom 200-ranked sentences from the Amazon data source.

Figure 6 compares each of the scoring methods to a random chance baseline (diagonal grey line), where the baseline indicated a number of pathway to injury sentences expected to be found in a random selection of that number of sentences on the holdout set taken from the Amazon data source. For each scoring method, Figure 6 illustrates the number of pathway to injury sentences (y-axis) found in the top-ranked N sentences using that scoring method. We created Receiver Operating Characteristics (ROC) curves, and calculated the area under curve (AUC) statistic (Bradley, 1997) to test the performance of each method. Generally, AUC scores getting close to 1 imply that the best models rarely contain false positives in the top-scored items as well as occasionally consist of false negatives in the bottom-scored items (Law et al., 2017). Additionally, Figure 6 showed graphically that the smoke term lists are more effective than sentiment analysis in finding the pathway to injury sentences over the whole of the distribution. Interestingly, while the trigram list was the most effective method in

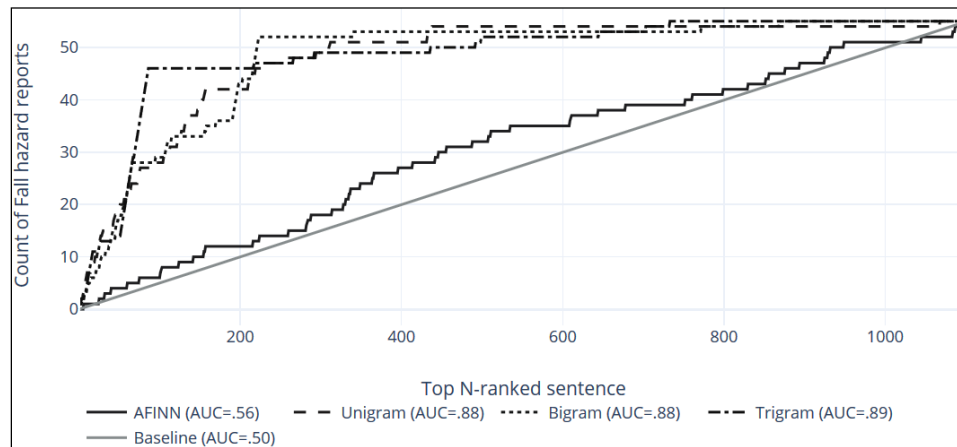
the top-ranked set of sentences, the unigram list was more effective over the entire distribution of sentences, measuring a superior AUC.

Figure 6. Pathway to injury-discovery performance of smoke scoring methods and sentiment approach (AFINN) on the holdout set of Amazon data source.



When analyzing a specific code value, e.g., fall from the hazard type, we assessed the performance of each smoke scoring method to a random chance baseline, where this baseline indicated a number of fall hazard sentences expected to be found in a random selection of that number of sentences. Figure 7 shows visually that the smoke term lists were more effective than the sentiment approach in discovering fall hazard sentences over the entire distribution. While the trigram list was the most effective method in counting 46 fall hazard sentences in the top 200-scored sentences, for instance, the trigram scoring method was also successful in fall hazard-discovery over the entire distribution of reviews, producing a superior AUC value.

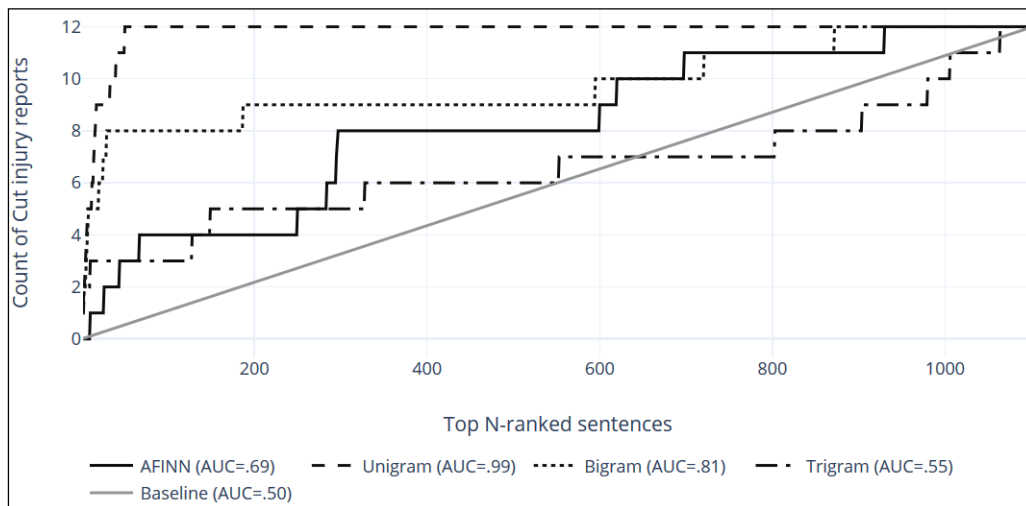
Figure 7. Fall hazard-discovery performance of smoke scoring methods and sentiment approach (AFINN) on the holdout set of Amazon data source.



To investigate a specific code value, e.g. cut injury from the injury category, we evaluated the performance of each smoke scoring method to a random chance baseline, where this baseline shows the number of cut injury sentences estimated to be found in a random selection of that number of sentences. Figure 8 indicated visually that the smoke term lists were more effective than the sentiment approach in identifying cut injury reports over the entire distribution. The unigram scoring method was the most successful in discovering cut injury over the entire distribution, producing a superior AUC value. Appendix D represents graphically the comparison of the scoring methods with the conventional sentiment method, AFINN, in detecting the rest of the code values across the top N-ranked sentences for the event, hazard, and injury types. Even though each of the sentiment methods was fairly effective in the top-scoring portion of sentences, each smoke term method still outperformed each sentiment method over the entire distribution of sentences as measured by the AUC scores (see Figures 12-17 under Appendix D). Each practitioner must decide as to the balance between maximizing true positives in the top-ranking reviews and considering the overall

AUC offered by different methods. Practitioners with limited resources pursuing rapid information from online reviews may only have a short time to read the top-ranking (e.g., top 50-ranked) set of sentences related to a code value (e.g., cut injury) of a specific safety concern category (e.g., injury type). For practitioners with more resources for category-specific safety surveillance, the higher AUC score may make smoke unigrams more desirable (see Figure 8).

Figure 8. Cut injury-discovery performance of smoke scoring methods and sentiment method (AFINN) on the holdout set from Amazon data source.



4.3. Generalizability of category-specific smoke n-grams

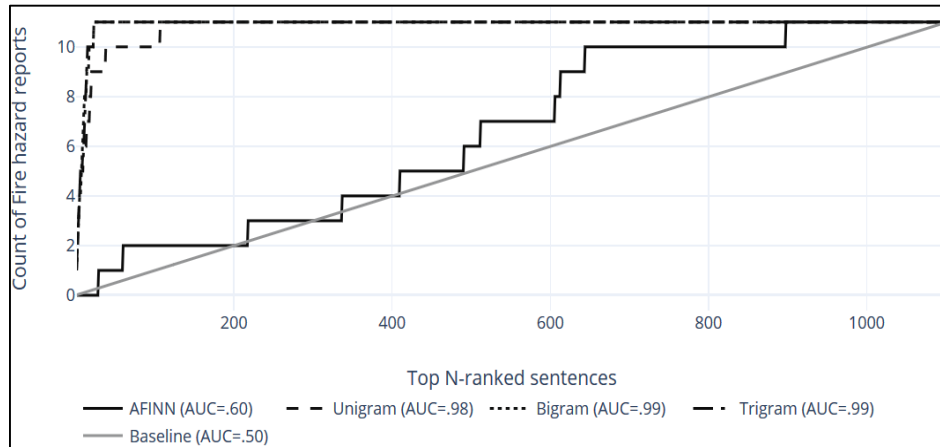
To evaluate the generalizability of our smoke term-scoring methodologies amongst various code values of safety concern categories, we test the methods on a further unseen dataset from a different online retailer. Questions of generalizability come into consideration when evaluating whether the smoke terms generated from Amazon’s training dataset, containing the safety concern reports of diverse product categories, also apply to a new dataset containing

the specific safety reports of different product categories. Hence, we assess the performance by comparing AUC scores of each smoke scoring method in detecting the reports of a code value of the category-specific safety concerns by using the new data source, i.e., safety concern reports collected from the Fortune 50 retailer based in the U.S.

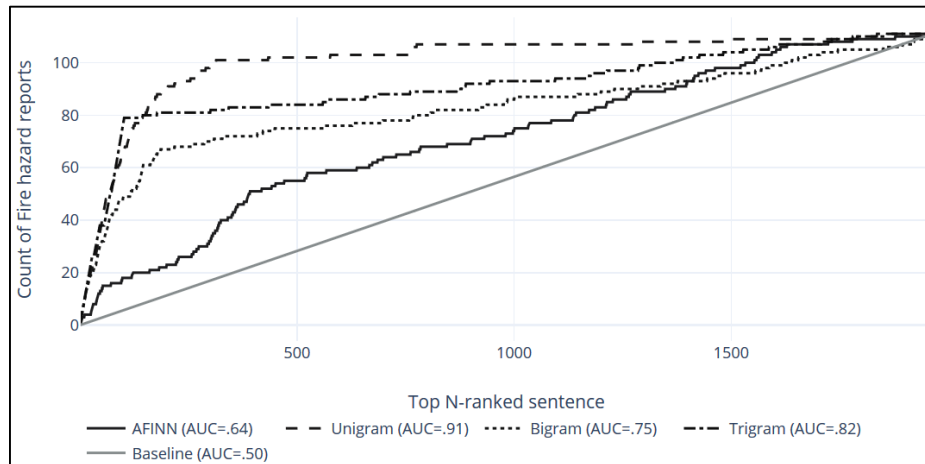
We assessed the smoke scoring method for detecting a code value of a hazard type, e.g., fire hazard to the whole dataset (i.e. 1,958 sentences) of the Fortune 50 retailer. Figure 9a showed graphically the distribution of the n-grams to discover the number of fire hazard reports among the top N-ranked sentences taken from the dataset of Amazon. When comparing the Figures 9a (Amazon) with 9b (Fortune 50 retailer), we observed a similar trend that the smoke term lists were more effective than the sentiment words in identifying the fire hazard reports over the entire distribution. Additionally, Table 12 under Appendix E shows the similar AUC scores of n-grams in identifying the fire hazard reports over the overall distribution. Hence, these results show the generalizability of the fire hazard smoke scoring method.

Figure 9. Fire hazard-discovery performance of smoke scoring methods and sentiment approach

(a) Amazon



(b) Fortune 50 retailer.

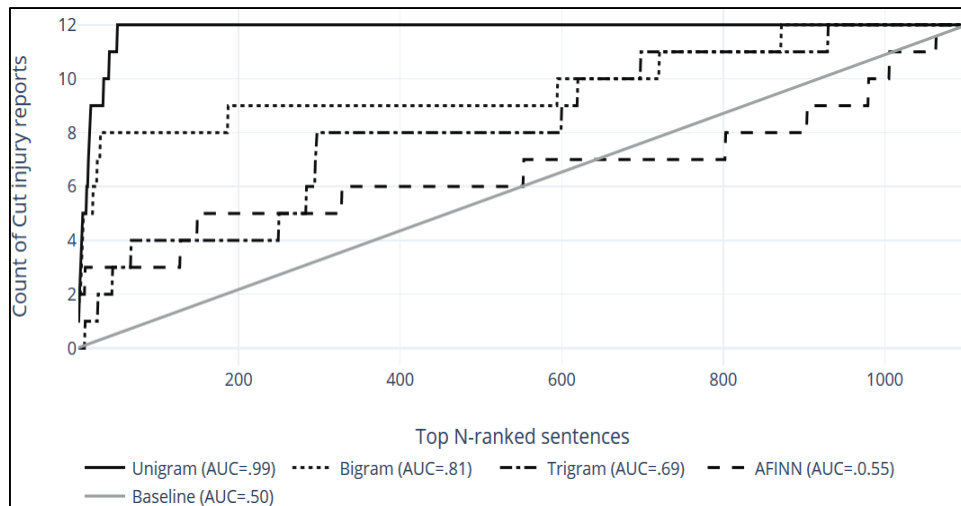


We then evaluated the generalizability of the smoke scoring method in detecting a specific code type of an injury category, e.g. cut injury by using the dataset of Fortune 50

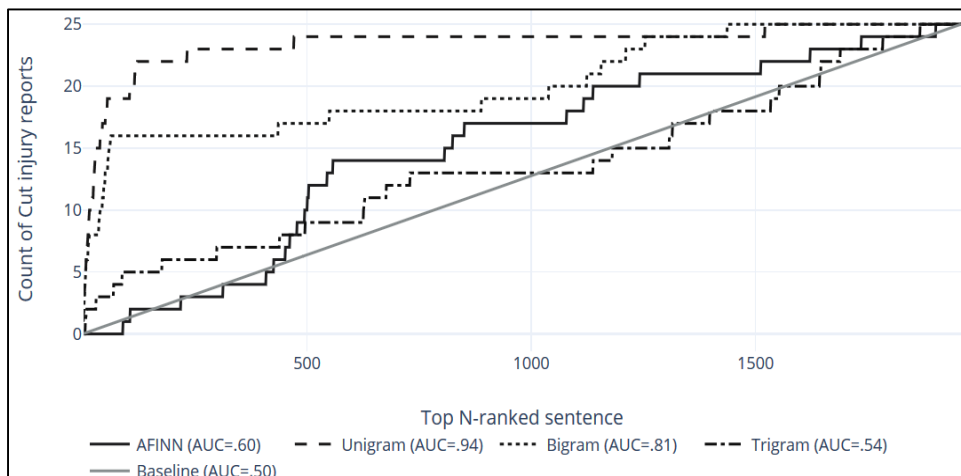
retailer. Figure 10a (Amazon) represents the overall distribution of the n-grams to identify the number of cut injury reports that extended across the top N-ranked sentences. As shown in both the Figures 10a (Amazon) and 10b (Fortune 50 retailer), there was also a similar trend of the smoke term lists being more effective than the sentiment word list in discovering the cut injury reports throughout the whole distribution. Additionally, Table 13 under Appendix E showed the similar AUC scores of n-grams. Therefore, the results indicate some ability of the smoke term scoring methods to generalize: we see similar AUC scores of each n-gram list in discovering the cut injury reports in the new dataset taken from the Fortune 50 retailer.

Figure 10. Cut injury-discovery performance of smoke scoring methods and sentiment approach

(a) Amazon



(b) Fortune 50 retailer.

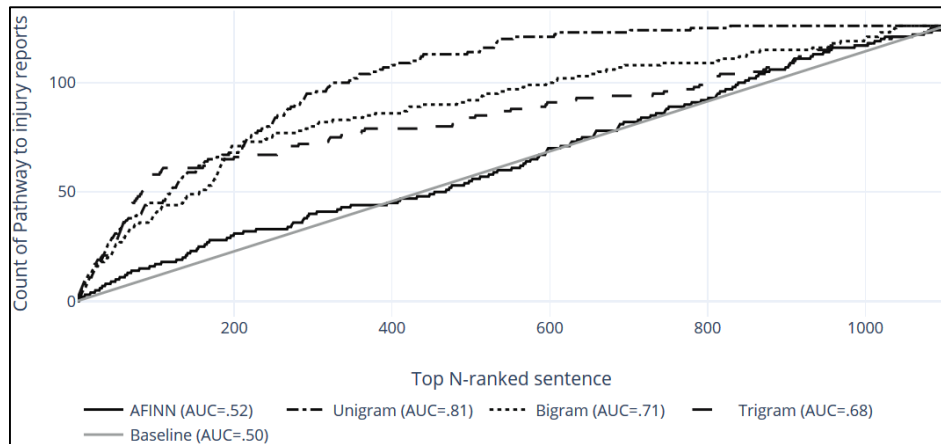


In contrast, Figure 11a (Amazon), Figure 11b (Fortune 50 retailer) showed a surprising result in that the smoke term method did not appear to generalize well. According to Table 14 under Appendix E, we found a lower AUC score for each n-gram when testing the performance in identifying the pathway to injury event on the dataset of Fortune 50 retailer. On the other hand, assessing the performance of each n-gram on the holdout set of Amazon

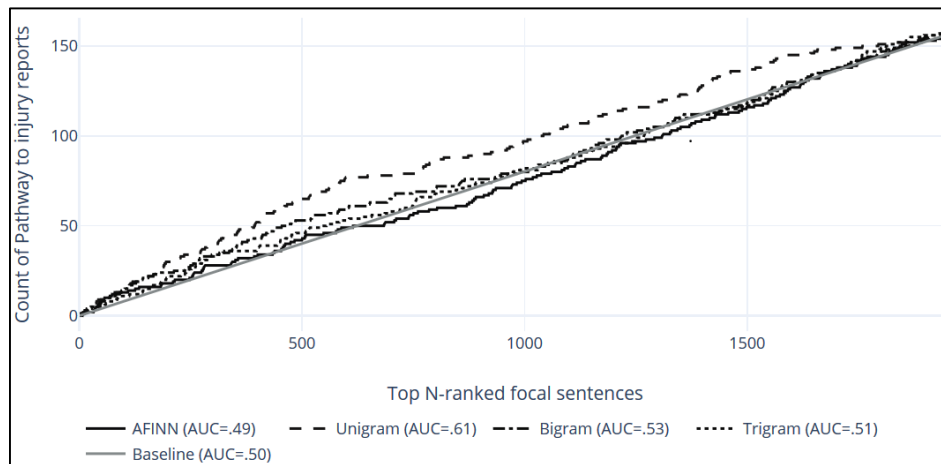
data revealed a high AUC score for each n-gram. Hence, the results (trained from the Amazon dataset) did not generalize the application of using the smoke term scoring methods in detecting the reports related to pathway to injury event on a dataset derived from a new data source (Fortune 50 retailer). Additionally, Appendix D showed similar charts that found low AUC scores for each n-gram list in counting the reports related to each code value of the event category, for the second retailer (Fortune 50 retailer).

Figure 11. Pathway to injury event-discovery performance of smoke scoring methods and sentiment approach

(a) Amazon data set.



(b) Fortune 50 retailer.



5. Conclusion

In this study, we used different text classification methods to segment the safety concern reports from an enormous volume of consumer-generated reviews across a diverse set of product categories. We observed that a small portion of our tagged dataset across multiple product categories actually referred to safety concerns. Our text analytics approach suggests that words are prevalent in safety concern reports within particular event, hazard, and injury categories. We used unsupervised topic modeling techniques to reveal six latent topics that were related to only hazard categories present in these reviews, but by using a supervised learning technique, we provided terms that practitioners and researchers could use to rapidly categorize the safety concern reports based upon event, hazard, and injury categories. Doing so adds specificity to safety surveillance and allows both firms and regulators to analyze online reviews for safety concern reports in greater detail. Additionally, we discovered that the sentiment approach was less effective than the supervised learning mechanism in predicting the category-specific safety concerns. Hence, the category-specific safety concern terms were able to isolate safety concerns more reliably. Finally, after assessing the supervised classification methods on one dataset (Amazon), we also used another dataset taken from a different source (i.e., the second Fortune 50 retailer) for evaluating the generalizability of these classification methods.

6. Limitations

Our study is confined to several limitations. The first is that we observed a large discrepancy in the numbers of reviews across product categories. Due to the broad coding of some product categories, there is also some potential bias in selecting event, hazard, and injury categories. For

example, toy products may consist of safety concern-related reviews with more choking hazards than fall hazards. Despite this possible selection bias, our inclusion of 17 different product categories does help to ensure the broad generalizability and applicability of our results. However, we found surprising results from assessing the supervised classification methods on the dataset of the Fortune 50 retailer because the product categories were different from the ones found in the Amazon dataset. A further limitation is that we utilized manual (human) tagging in order to determine which reviews actually referred to category-specific safety concerns. Each assessment of whether a focal sentence referred to a specific safety concern is subjective, inducing some variability from tagger to tagger. To handle this limitation, we first ensured that focal sentences along with their corresponding reviews were distributed to taggers at random. Due to random assignment, the taggers A and B each assessed some sentences, ensuring that several perspectives were considered. Additionally, we performed a final round of tagging in which an “authority” tagger verified students’ tags.

7. Implications

Our findings have important implications for product safety experts and also for researchers.

7.1. Implication for practice

In terms of implication for the practitioners, we found that online reviews are a substantial source of safety concern-related feedback across multiple product categories. However, we also found substantial variability across product categories in the proportion of reviews that referred to category-specific safety concerns. Due to this reason, sentiment

analysis cannot be relied upon to monitor particular safety concerns effectively in most product categories. While sentiment analysis continues to be a valuable tool for detecting the customer's emotive opinions about products, it is not the most efficient way for the practitioners to identify category-specific safety concerns across multiple product categories. Our results prove that the smoke-scoring methodology discovers substantially more non-emotive opinions about products than sentiment analysis. These non-emotive narratives are more prevalent in category-specific safety concerns than emotive instances. This study highlights that the safety concern-specific terms derived from safety concern sub-categories retrieve more actionable insights of product safety than the non-specific emotion-focused terms supplied by traditional sentiment analysis. Therefore, by using our effective safety concern-specific smoke term dictionaries (generated in this study), the practitioners such as product safety experts should take advantage of rapidly identifying, categorizing as well as prioritizing the specific sub-categories of safety concern reports from online reviews to allow rigorous evaluation of safety concern sub-classification. Our research has important policy implications: regulators should move rapidly to establish a clearinghouse of hazard reports from online reviews from multiple retailers in order to allow more effective smoke dictionaries to be trained, and to allow more rigorous evaluation of hazard identification and hazard-sub-classification machine learning classifiers.

7.2. Implication for research

Next, we demonstrated in our study that, although different product categories have distinct safety concerns, there are some commonalities in hazard categories across product categories. These commonalities may be difficult to observe in smaller dataset, but using a large dataset of over 120,000 reviews, we were able to collect enough verified safety concern reports to

reveal a wide range of hazard categories by using the supervised classification method. In terms of research implications, it is clear that future research needs to gather substantially larger datasets for certain safety concern types. Specifically, we need more reports of the low-incident target classes (Event: injury mitigation, treatment; Hazard: biological; Injury: electric shock, foreign bodies, fracture, head injury, poisoning), so that supervised learning for these code values can proceed with sufficient positive instances for training and testing.

8. Future work

Future work should develop more robust assessments of the safety concern-specific smoke term lists, especially for the low-incident target classes.

In conclusion, our results provide sets of unique terms that are useful in identifying, categorizing and prioritizing category-specific safety concerns reports across a broad spectrum of product categories. Specifically, we demonstrated that safety surveillance can be performed on a more granular scale than considered in prior work. We showed this facet through our observed safety concern categories.

In the next chapter (Chapter 2), we will utilize some of the safety concern sub-classification techniques defined in this chapter (Chapter 1) to develop a novel framework for risk assessment of various products. Specifically, we will leverage the hazard type sub-classification techniques, to allow organizations to determine risk scores of products on the basis of this further level of categorization. Our aim is to construct a risk assessment model that disentangles multiple categories of safety concern reports, using hazard type sub-classification and uses risk assessment criteria to prioritize the investigation of products with the most serious safety concerns.

References

- Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6), 975-990.
- Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems*, 55(4), 871-882.
- Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, 54(1), 87-97.
- Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International journal of medical informatics*, 100, 108-120.
- Alemzadeh, H., Iyer, R. K., Kalbarczyk, Z., & Raman, J. (2013). Analysis of safety-critical computer failures in medical devices. *IEEE Security & Privacy*, 11(4), 14-26.
- Allen, L.P., Comolli, R., & Heumann, S. (2008). China product recalls: what's at stake and what's next. NERA Economic Consulting.
- Bapuji, H., & Beamish, P. (2008). Product recalls: avoid hazardous design flaws. *Harvard Business Review* 86 (3), 23-26.
- Boutros, B. A. S., Hewedi, M. M., Roberts, K. R., & Megahid, F. M. (2014). Food safety traceability systems in the maritime catering logistics. *Food and Nutrition Sciences*, 5(15), 1447.
- Brynjolfsson, E., Hu, Y., & Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49(11), 1580-1596.
- Health Canada. (2016). *Injury Coding Manual*.
- CPSC. (2009). Maclaren USA recalls to repair strollers following fingertip amputations. Retrieved from <http://www.cpsc.gov/cpscpub/prerel/prhtml10/10033.html>.
- Reporting Product-Related Injuries and Deaths to the U.S. Consumer Product Safety Commission (CPSC). (2014). *AAEM NEWS*. Retrieved from <https://www.aaem.org/UserFiles/SepOct14CPSC.pdf>
- Dhar, V., & Chang, E. A. (2009). Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing*, 23(4), 300-307.

- Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems*, 40(2), 213-233.
- Gallozzi, M., Tucker, S. (2007). Insuring Against Disaster: Coverage for Product Recalls, The Insurance Coverage Law Bulletin, October, 2007
- Goldberg, D. M., & Abrahams, A. S. (2018). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*, 105, 52-65.
- Gregg, R. A. (2019). *Protecting Americans from Dangerous Products: Is the Consumer Product Safety Commission Fulfilling its Mission*. Retrieved from <https://docs.house.gov/meetings/IF/IF17/20190409/109318/HHRG-116-IF17-Wstate-GreggR-20190409.pdf>
- Hora, M., Bapuji, H., & Roth, A. V. (2011). 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), 766-777.
- Jaimes, A., Sebe, N., & Gatica-Perez, D. (2006, October). Human-centered computing: a multimedia perspective. In *Proceedings of the 14th ACM international conference on Multimedia* (pp. 855-864). ACM.
- Kanayama, H., & Nasukawa, T. (2012). Unsupervised lexicon induction for clause-level detection of evaluations. *Natural Language Engineering*, 18(1), 83-107.
- Kuiper, P. K., & Hood, K. M. (2019, January). Examining Sentiment Analysis When Evaluating Survey Responses. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* (pp. 412-415). IEEE.
- Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics*. 33 (1), 159–174.
- Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67, 84-94.
- Liu, Y., Feng, J., & Liao, X. (2017). When Online Reviews Meet Sales Volume Information: Is More or Accurate Information Always Better?. *Information Systems Research*, 28(4), 723-743.
- Marucci-Wellman, H., Lehto, M., & Corns, H. (2011). A combined Fuzzy and Naïve Bayesian strategy can be used to assign event codes to injury narratives. *Injury Prevention*, 17(6), 407-414.
- Marucci-Wellman, H. R., Corns, H. L., & Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review. *Accident Analysis & Prevention*, 98, 359-371.

- McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015, August). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 43-52). ACM.
- Mummalaneni, V., Gruss, R., Goldberg, D. M., Ehsani, J. P., & Abrahams, A. S. (2018). Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety science*, 104, 260-268.
- Nanda, G., Grattan, K. M., Chu, M. T., Davis, L. K., & Lehto, M. R. (2016). Bayesian decision support for coding occupational injury data. *Journal of safety research*, 57, 71-82.
- Nasri, L., Baghersad, M., Gruss, R., Marucchi, N. S. W., Abrahams, A. S., & Ehsani, J. P. (2018). An investigation into online videos as a source of safety hazard reports. *Journal of Safety Research*.
- Pan, S., Wang, L., Wang, K., Bi, Z., Shan, S., & Xu, B. (2014). A Knowledge Engineering Framework for Identifying Key Impact Factors from Safety-Related Accident Cases. *Systems Research and Behavioral Science*, 31(3), 383-397.
- Poston, A. (2018). *Guidance on the Application of Human Factors to Consumer Products*. Retrieved from <https://www.cpsc.gov/s3fs-public/HF-Standard-Practice-Draft-12Feb2018.pdf?CGk4Zs9GabjCnZ5RXQuSlr2toQ1aLPhJ>
- Shaaban, K., Maher, A., & Singhapakdi, A. (2019). Effect of vehicle safety recalls on the perception towards other vehicles from the same country of origin. *Journal of Traffic and Transportation Engineering (English Edition)*.
- Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems. *Journal of Management Information Systems*, 34(4), 1054-1081.
- Trottman, M., & Mitchell, J. (2010). Toyota recall delay comes under scrutiny. *Wall Street Journal*, January 29, 2010, accessed on: January 30, 2010.
- Vallmuur, K., Marucci-Wellman, H. R., Taylor, J. A., Lehto, M., Corns, H. L., & Smith, G. S. (2016). Harnessing information from injury narratives in the 'big data' era: understanding and applying machine learning for injury surveillance. *Injury prevention*, 22(Suppl 1), i34-i42.
- Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision support systems*, 90, 23-32.
- Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58, 51-65.
- Yao, L., & Parlar, M. (2019). Product recall timing optimization using dynamic programming. *International Journal of Production Economics*, 210, 1-14.

Zhang, X., Qiao, Z., Ahuja, A., Fan, W., Fox, E. A., & Reddy, C. K. (2019). Discovering Products Defects and Solutions from Online User Generated User Generated Contents. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA. <https://doi.org/10.1145/3308558.331373>

Appendix A. Tagging protocol of the code values for each event, hazard, and injury code category

Tagging instructions:

Each review contains a specific *sentence* that will indicate for your tagging. |||= This sentence is formatted as a focal sentence with vertical bars and equals signs on either side. ==|| Please only focus on the discussion that is indicated in the focal sentence. The entire review is included with the focal sentence for giving it a proper context, but please tag only based upon the discussion that is present in the focal sentence. If the sentence discusses multiple types of safety defects, then please check each corresponding box. If the sentence does not mention any one of the types, then please check the box labeled “none”.

- **Tag Code types of Event Category-** Pathway to Injury Event/ Potential Injury Event/ Actual Injury Event/ Treatment Event/ Injury Mitigation Suggestion /Product Remediation Event/ None

Pathway to Injury Event: occurs when a sentence mentions non-human subject or non-human object and an injury can occur after the event (e.g. house caught fire, shelf tipped over, fall from swing).

“I purchased this table a couple weeks ago, not only was the assembly difficult but the corners of the table are extremely sharp.”

*“The blades can't handle any impact with a hard surface, **they tend to break off into small sharp flying projectiles.**”*

*“This **belt leaked gasoline all over the lawn** more engine I returned it to the store.”*

*“caught on fire.... a few months of use, one day I just happened to be in the garage I was warming up and **it just started to flame all around the back of the cone** where the wire was.”*

*“Had I not been in there at time **it probably would have caught my barn on fire** (flame was going straight up several feet).”*

*“We came home with our son and my husband sat in the chair and the **legs just busted and broke.**”*

Potential Injury Event: occurs when an injury verb is used (denoting pain or damage to a person), but with a qualifier that it may happen. The affected item is a human or human body part, and the event describes that the human or body part may be hurt or damaged (e.g. could burn my arm, could crush his skull, may get burned)

*“We never use it because even a grown **adult can fall right through it**”*

*“I was afraid **someone would get a hand or a head stuck as the rest of their body went sliding past**”*

*“...because with the guard being minuscule, and while using as an edger, everything, and I mean **everything, flies up in your eyes, face, hair and body.**”*

*“Canvas is flimsy -will last only a short time. Umbrella held by a pin - be prepared to catch umbrella or **it will come down on your head.**”*

*“They flicker so much **I fear I’ll have a seizure.**”*

Actual Injury Event: occurs when an injury already occurred, e.g. “One time she fell into it she got a huge knot on her forehead w/a nasty purple bruise”.

*“I **broke my arm and my brother almost died** from landing on his back.”* This clause indicates a presence of Actual Injury event: “broke my arm”.

*“I have **sliced my finger open twice** while cleaning the inside of this bird feeder.”*

*“**Burn your arm**; carryover tube is now at front. There is **no hair left on my arm** just behind the wrist”*

Treatment Event: occurs after a human is injured by the product and then receives some medical remedy an injury that already occurred (e.g. “gave ten stitches”).

*“As soon as it was released, the blades started moving and cut the **palm of my hand requiring a trip to the ER and stitches**”*

*“The **doctor removed a piece from his lung**”.* This clause indicates a remedy taken place to treat an injury.

Injury Mitigation Suggestion: occurs when the user suggests some possible remedy for the product manufacturer to reduce the risk of product-related injury (e.g. “should make it safer by removing the sharp edges”)

*“This **table should be built with tempered or a stronger type of glass**”*

*“And now, about the suckered side, if you hold the blower with your left hand, **wear a thick shirt.**”*

*“**if they improve the connection and get rid of the gap** it could possibly be a decent toy”.* This clause indicates a suggested future fix for a potentially unsafe product.

Product Remediation Event: occurs when the user takes some action to reduce the risk of injury before or after an actual injury (e.g. “padded all the sides”) by upgrading the product for making it safe to use.

*“I **patched these up with pieces of aluminum and hose clamps** until I can replace this unit... hopefully before it burns down my home.”*

*“The machine can be dangerous, **always keep tight grip and also I use up and down motion to prevent getting stuck in roots.**”*

*“Luckily we had **put a foam-like interlocking pad squares in the play room** and we had just figured we'd put the slide on top of those pads.”* This clause indicates an action taken place to reduce a potential injury.

None Event: There is not enough information to tag the sentence as any one of the event types (listed above).

“After assembling the walker I received, I noticed that it listed to one side.”

“I literally had to throw this away the other day after using it....”

“They are sharp!”

The sentences mentioned above does not give out enough information to tag one specific event type.

- **Tag Code types of Hazard category-** Choking/ Chemical/ Biological/ Electrical/ Fire/ Fall/ Hitting/ Mechanical/ Strangulation/ Suffocation/ Swallowing/ Thermal/ Other/ None

Choking hazards: occur from objects “such as small parts that come loose from products, e.g. toys, which may be partially swallowed by a child and make the child choke”⁵ when the object blocks the respiratory tract, causing the child hard to breath.

*“I had to rip off all the stickers because I was afraid **they would peel them off and choke on them.**”*

*“I caught my **5 month old baby with the entire triangle portion of the toy in her mouth**”*

*“There is always the potential that **they could chew the cardboard off and choke** but as always, supervision is the key here”*

Chemical hazards: occur from the “improper use of pesticides or antimicrobial residues, toxic chemicals used on processing equipment such as oils used on equipment or sanitizers.

Furthermore, other potential chemical hazards (e.g. toxic smoke or gas) may include substances that are used in processing at certain levels but can cause illness or difficulty in breathing if consumed/inhaled at too high of a concentration, such as sodium nitrite or antimicrobial solutions,”⁶ or a substance could cause injury if touched on skin.

*“This is a great toy other than the fact **it uses a toxic chemical**”*

⁵ <http://ec.europa.eu/>

⁶ <https://www.canr.msu.edu/>

*“I will now have to choose between leaks or **my child getting another chemical burn and suffering.**”*

Biological hazards: “include microorganisms such as bacteria, viruses, yeasts, molds and parasites. Some of these are pathogens or may produce toxins. A pathogenic microorganism causes disease and can vary in the degree of severity. Examples of biological hazards include *Salmonella*, *E. coli* and *Clostridium botulinum*.”⁷

Fire hazards: occur from electrical systems (e.g., heater fan) that overheats, causes smoke and/or burning smell, and/or catches fire. May also catch fire due to a flammable gas or oil (e.g. propane, kerosene, etc.) leak.

*“**the plug started smoking** heated up to such an extent that it melted the plug on my extension cord as well as it's own plug”*

*“When I went to lay in bed **I could smell something electrical burning.**”*

*“The **second one leaked propane around the tube on the heater** where it threads into the tank.”*

Fall hazards: occur from unintended loss of balance of any human body part, e.g., “when he started to stand up and cruise, he would fall often (as any baby does) and more than twice he fell...”

*“Seat too high for my toilet; Cannot secure it firmly; Unless you use equal pressure on both side arms at the same time **it tips and throws you off balance.**”*

*“Once it clicked I figured I'd done something wrong because **it was SO wobbly.**”*

⁷ <https://www.canr.msu.edu/outreach/>

*“It was used in a chair without arms and **she was able to lean over way too far**”*

Hitting hazards: include collision with an object, e.g. “he threw the smallest bucket at my face and it hit so hard that now I have a black eye”.

*“The idea of taping a motor on is horrible, and I am sure if it hasn't yet **it will eventually fall off and land on a child who will lose an eye**”*

*“**LOVE*D*** this hammock but yesterday **it broke with my hubby on it**”*

*“Be warned that when it hits something unloveable, **it will twist and can knock an unsuspecting user over**”*

*“if they lunge forward when the tray isn't there **they will almost flip themselves out- head first**”*

Fall and Hitting hazards: include a person falling would also hit something afterward.

*“Next thing you know the **entire stroller with baby tipped over.**”* This sentence shows a fall hazard as the baby’s stroller is leaned over. Additionally, there is also a potential hit hazard the baby would hit the floor afterward.

Mechanical hazards: result from either powered or manual (human) use of tools or equipment, e.g., moving or rotating parts/object/product); “hard or sharp edges such as broken glass, metal, plastic, stones, pits, wood, or even bone”⁸ that “can cut fingers, or tight openings in which someone can trap their fingers”⁹.

*“However the build quality is not great - some **sharp edges to cut yourself on.**”*

⁸ <https://www.canr.msu.edu/outreach/>

⁹ <http://ec.europa.eu/>

*“You pull the trigger, your hand is at the top and the **sharp plastic edge cuts your and just like a paper cut.**”*

*“As we were putting in the umbrella, the **tabletop glass shattered.**”*

*“Bought this **glass and it SHATTERED** with the umbrella with a little wind.”*

Strangulation hazards: result from squeezing the vein or any body part tightly by external object (e.g. cord, rope, etc.) so there is a blockage of blood circulation or airflow, e.g. a baby could get her head or hand caught between the gaps of a baby crib or a blind cord could get wrapped around the neck causing lack of airflow.

*“i was trimming along nicely, and all of the sudden the **handles would start to twist in my hand**”*

*“The adjustable height is nice but **arms get CAUGHT in between the seat and the arms**”*

Suffocation hazards: occur during the restriction of airflow through windpipe, e.g., baby’s head flopping forward that does not allow the baby to breathe properly, or this hazard could occur due to an object covering the face or nose causing a difficulty in breathing.

*“headrest causes head to lean forward - my **son's chin almost touching his chest, not good for breathing, may cut off air passage**”*

*“The mesh is so fine that **it makes it hard to breathe**”*

Swallowing/ingesting hazards: result from swallowing and ingesting a small object (e.g. magnet or a broken piece), and then getting into the digestive system, e.g. stomach or intestine.

*“You definitely don't want loose bolts within reach of small children, and **these are just the right size where I could see a child trying to swallow one**”*

*“they are packaged in this cheap plastic baby bottle and **they really look like they should be eaten (they are far too small to be of use as a functional eraser)**”*

*“**she swallowed a screw and was taked to the hospital** for x-rays and surgery.”*

Thermal hazards: result from hot outer surface of an oven (i.e., high temperature components) that can lead to skin burns.

*“My husband nearly **Burned his fingers** when trying to lift the lid”*

*“Well about half way into the day I had to switch back to the old gloves as **these were causing my hands to burn**, now I have blisters and red hands”*

*“The actual knob isn't hot but **it's fairly easy to burn your hands** turning off the gas line if you're not super careful”*

Other: a specific hazard type that does not describe any one of the above lists.

None: does not describe a specific hazard type.

“i also hated the sharp corners.”

“I bought the unit and it was working very well.”

- **Tag Code types of Injury Category- Allergic Reaction/ Asphyxia/ Brain Injury/ Bruise/ Burns/ Cut/ Electric Shock/ Foreign Bodies/ Fracture/ Poisoning/ Other/ None**

Injury Type	Definition	Description
Allergic Reaction	“antibodies traveling to cells that release chemicals, causing symptoms most often in the nose, lungs, throat, sinuses, ears, lining of the stomach or on the skin” ¹⁰	Result of rash, burns, irritation, etc. due to: Corrosive substances that contact the skin; General chemical products; General cosmetic ingredients; Toilet bowl cleaners; drain cleaners, ammonia, bleaches, rust removers, deodorizers, sanitizers, paint removers, disinfectants, etc.
Asphyxia	“classify a category of injuries that pose a threat to breathing or result in oxygen deficiency” ¹⁰	Absence of oxygen due to: obstructing air passage through posture or chest compression; lodging object in mouth or airway; sealing the mouth and/or nose with external object (e.g. bag over the head, blanket over the mouth and nose); giving external pressure (e.g. blind cord wrapped around the neck);
Head Injury	“caused by a bump, blow or jolt to the head only” ¹⁰	Headache, dizziness, feeling unsteady; Nausea;
Bruise	“A bruise is an area of skin discoloration. A bruise occurs when small blood vessels break and leak their contents into the soft tissue beneath the skin” ¹⁰	Pain; swelling; tenderness; confined to small area; bleeding under the skin and excess fluid may accumulate; Pinching that does not cause a break in the skin; Friction burn that does not cause a break in the skin;
Burns	“A burn is an injury to the skin or other organic tissue primarily caused by heat or due to radiation, radioactivity, electricity, friction” ¹⁰	Arch or flash burns caused by an electric arc or explosion; Hot liquid or steam (scalds); Hot objects (contact with hot surfaces of appliances, overheated electric conductors or equipment; clothing or textiles catching on fire due to electrical malfunction or flammability hazard); Open fire, flames, coals; General consumer

¹⁰Health Canada. (2016). *Injury Coding Manual*

		product examples: Laser pointers causing burns to eyes, Welding flash;
Cut	“A break in the skin or mucous membrane”; “classify injuries that are also referred to as open wounds” ¹⁰	Penetrate by a sharp edge or object; Irregular or jagged or blunt breaking or tearing away of the skin by often associated with tools and machinery; piercing wound by a pointed object (splinter, nail, needle, dart); scraping off topmost layer of the skin
Electric Shock	“physiological reaction to electricity flowing between parts of the body or through the body to a ground” ¹⁰	Human body has made contact with an electrical source through the entry or exit of the electrical current, but there may not be a burn or a skin discoloration

Foreign Bodies	“category encompasses both foreign bodies that have been swallowed that may or may not pass through the gastro-intestinal tract as well as foreign objects entering the body through the ear, nose, eye, genitals or rectum.” ¹⁰	Swallowed objects: foreign bodies that pass through the esophagus and into the stomach without getting stuck will probably pass through the entire GI tract.
Fracture	“A fracture is a broken bone. A bone may be completely fractured or partially fractured in any number of ways (crosswise, lengthwise, in multiple pieces)” ¹⁰	Crushing injuries including one or more fractured bones; Injuries referred to as ‘broken bones’; Fractures with the bone breaking skin; Chipped or broken tooth
Poisoning	“A toxic exposure occurs when a susceptible individual or population comes into contact with a chemical in a particular environment. The individual may be	Biting into or eating a liquid pod pack; Lead/heavy metal poisoning; Carbon Monoxide poisoning;

	poisoned as a result” ¹⁰	
Other	An injury is mentioned, but not one of the types listed above	-
None	There is no injury mentioned	-

Examples of each Injury type

Allergic Reaction:

*“children got a bad **rash** or **chemical burn** from the Dry Max diapers, but I decided to try them anyway.”*

*“I had an **allergic reaction** to washing my laundry with soap nuts.”*

Asphyxia:

*“child was actually choking and **not breathing**...”*

*“daughter nearly **choked to death** today on the top red piece on the Rainbow Tower”*

Head Injury:

*“I **hit** the back of **my head** and sprained my wrist in the shower.”*

*“I just bought another chair after my father fell over backwards on this chair and almost **cracked his head**”*

Bruise:

*“I took off her PJs to discover **both legs** were **dark purple**.”*

*“The snap in the center was also so tight on her **little legs** (it says in the instruction manual to only place the latch in the first or second slot for rear facing) it **left marks** on her.”*

Burns:

*“I can't bend one of **my fingers** because my **knuckle got burnt** so badly on the *outside* of the glass door.”*

*“The entire case and glass door gets **so hot** during use that it **burns skin**.”*

Cut:

*“Her left hand accidentally touched the wide and long feather touch switch that set the blades rotating while her **right index finger** was **still caught in** the cup of her blender.”*

*“When she did wear them (for about an hour) she found that her **ankel hurt** because there was a inside part of the boot **protruding** into **her ankle**.”*

Electric shock:

*“I have gotten a few good **shocks** from things containing metal that rested near it.”*

Foreign bodies:

*“Then she **swallowed a screw** and was taked to the hospital for x-rays and surgery. Don't buy their beds.”*

*“An autopsy found two **tiny magnets** stuck in Kenny's **small intestine**.”*

Fracture:

“I am afraid the fall down and his hand get stuck and break his hand.”

"I fell off and landed on my arm and my brother landed on his back. We went to the hospital and I broke my arm and my brother almost died from landing on his back"

Poisoning:

*"daughter was **lead poisoned** and had to be taken to the hospital."*

*"it gets warm but the **smell of gas** is enough to **knock you** out"*

Other:

*"Will not lay completely flat, cause **strain** on your mid back."*

*"unplugged the heater for the rest of the night and opened up my windows when it was 34 degrees outside to let fresh air in my room and my bedroom temp slowly went back to normal temp again and my sick feeling **overheated body** went back to normal later that night"*

*"The trigger does NOT have a locking mode like others I have used in the past your hand will get very **sore/strained** after 15 or so minutes of use."*

None:

"Then before I was finished mowing the belt had broke and mower was smoking very badly."

"Never had problems like this with any other gas powered equipment."

Appendix B. Smoke word lists for each code value under event, hazard, and injury categories

Table 6. Smoke word lists for each code value under event category

Pathway to injury event		
Word	Phrases	
	2-word	3-word
hazard	choking hazard	side to side
side	weight is	loose and the
choking	side to	a choking hazard
loose	the seat	secure it to
fall	to side	is applied to
toxic	to secure	the straps dont
wobbles	the straps	fall off and
plastic	off easily	come off easily
collapsed	containing lead	off within a
unstable	sharp edges	it caught fire

Potential injury event		
Word	Phrases	
	2-word	3-word
head	his head	fall out of
forward	her head	legs stuck in
neck	to choke	when she fell
hit	her neck	a sharp edge
mouth	legs stuck	the seat back
asleep	his chest	pull it out
chin	cut your	loud crack and
swallowed	to death	my sons chin
hurt	slid off	seat back a
face	hit the	causes head to

Actual injury event		
Word	Phrases	
	2-word	3-word
rash	my skin	to the hospital
blood	cut in	the edges of
skin	cut my	i cut myself
cut	edges of	jammed down his
gash	chemical burn	a chemical burn
accidental	the edges	edges of scissors
bleeding	the hospital	about the edges
hospital	emergency room	which needed stitches
cuts	marks on	razor sharp knives
sick	rash and	gash which needed

Treatment event		
Word	Phrases	
	2-word	3-word
saved	heal enough	while her right
patched	blender fortunately	left the hospital
ventilator	life after	after heal enough
index	finding the	caught in an
feather	5 stitches	accidently touched the
performing	she left	stay with me
finding	then removed	performing emergency surgery
intestine	the er	minutes away from
lungs	and throat	finger was still
rotating	doctors patched	to start walking

Product remediation event		
Word	Phrases	
	2-word	3-word
pillows	but duct	but duct tape
detergent	towel behind	the fabric cover
towel	ways to	pin to secure
duct	before putting	a safety pin
tape	safety pin	use a safety
uninstalled	pin to	secure the fabric
bristled	a towel	safety pin to
soften	heavy duty	months i padded
riser	fabric cover	bottom to help
chalkboard	duct tape	lotion on her

Injury mitigation event		
Word	Phrases	
	2-word	3-word
swaddle	area of	area of the
needs	allow for	get some system
custom	the swaddle	support rails
fewer	padding and	can protect the
rubberized	stabilization in	way we cant
bolted	inspire confidence	removes this rebound
installs	magnets however	a wall it
flap	and shimmy	you can wipe
learning	and hurting	padding and that
workouts	product on	on needs legs

Table 7. Smoke word lists for each code value under hazard category

Fall hazard			Hitting hazard			Chemical hazard		
Word	Phrases		Word	Phrases		Word	Phrases	
	2-word	3-word		2-word	3-word		2-word	3-word
fell	she fell	fell out of	hit	hit the	in the head	chemicals	a chemical	defects or other
fall	almost fell	right out of	head	head on	head on the	toxic	to cause	or other reproductive
fallen	have fallen	would have fallen	smack	it hit	his head when	chemical	other reproductive	birth defects or
slippery	from falling	fall out of	hitting	hit his	in the water	smell	reproductive harm	known to the
chair	and fell	and fell off	collapsed	my face	it hit the	rash	rash and	the state of
falling	a fall	and she fell	flip	head when	smack his head	skin	birth defects	to the state
slippage	out of	i almost fell	face	to smack	onto the ground	irritation	the state	warning this product
handles	fell off	daughter fell out	crash	head into	hit the floor	reproductive	chemical burn	cause cancer and
weight	off the	from falling off	layout	is heavy	hit his head	fumes	chemicals that	to cause cancer
slip	the chair	if you lean	radius	smack his	in a crash	toxicity	exposed to	a chemical burn

Choking hazard			Fire hazard			Mechanical hazard		
Word	Phrases		Word	Phrases		Word	Phrases	
	2-word	3-word		2-word	3-word		2-word	3-word
choking	choking hazard	a choking hazard	fire	caught fire	it caught fire	cut	cut in	a sharp edge
choke	a choking	choking hazard so	smoke	on fire	fire in my	sharp	sharp edges	the edges of
hazards	choke on	pose a choking	burning	fire and	motor of the	edges	cut my	i cut myself
throat	hazard a	jammed down his	burned	the smoke	was home and	blade	i cut	pouch that takes
choked	choke and	small child and	caught	a fire	to turn off	stitches	the blade	edges of scissors
chews	to choke	and choke on	smoking	smoke and	burnt plastic smell	cutting	into your	that dig cut
chewing	pose a	a potential choking	sparks	fire after	of sparks a	gash	a deep	about the edges
pose	jammed down	choke on them	catching	fire in	the motor of	finger	edges of	which needed stitches
posing	down his	down his throat	burnt	it caught	of the dehydrator	cuts	cut myself	razor sharp knives
hub	throat and	it and started	melted	the motor	the circuit board	blood	edges to	by having edges

Strangulation hazard			Suffocation hazard			Swallowing hazard		
Word	Phrases		Word	Phrases		Word	Phrases	
	2-word	3-word		2-word	3-word		2-word	3-word
stuck	legs stuck	legs stuck in	head	his chest	my sons chin	swallowed	chunk of	chunk of lead
dangled	stuck in	either end stand	sided	head to	causes head to	magnets	the magnets	a chunk of
closing	his arm	arms will be	flopped	falls asleep	instructions headrest causes	intestine	a chunk	to the hospital
linings	in between	strap when carrying	suffocation	air passage	prop her head	chunk	was lead	the hospital along
rag	that dangerous	rails between the	mesh	check buckle	to use towels	swallowing	magnets from	lead poisoned and
bruises	unattended in	it comes unhooked	passage	towels to	back a little	penny	small intestine	right size where
neck	drawers to	arm back out	flopping	color height	buckle sometimes difficult	transient	tiny magnets	within reach of
rails	bruises easily	dangled like a	chin	head flopped	good for breathing	glance	magnets however	the doctors discovered
wrap	never leave	holes of the	forward	height weight	air passage check	autopsy	doesnt loose	the o ring
spinning	that neck	between this and	asleep	the mesh	use towels to	blenders	on board	rubber in her

Thermal hazard		
Word	Phrases	
	2-word	3-word
hot	glass door	i pick it
melamine	so hot	kernels actually pop
flooring	will damage	the vinyl flooring
uncooked	burning ourselves	the kernels actually
knuckle	but inside	flooring over time
barefoot	a miracle	of the glass
cabinet	and glass	too hot and
yellow	kernels at	gets really hot
shoots	oven thermometer	seem to burn
ripping	it barefoot	groceries on top

Table 8. Smoke word lists for each code value under injury category

Cut injury			Allergy injury			Asphyxia injury		
Word	Phrases		Word	Phrases		Word	Phrases	
	2-word	3-word		2-word	3-word		2-word	3-word
cut	cut in	a sharp edge	allergic	allergic to	blisters i dont	sided	his chest	my sons chin
sharp	cut my	the edges of	asthma	chemicals they	choose between leaks	passage	more wobbly	seat back a
edges	i cut	i cut myself	rash	the flame	the fragrance from	chin	air passage	causes head to
stitches	into your	pouch that takes	allergies	carseat my	or my child	throat	check buckle	instructions headrest causes
cutting	a deep	too small for	skin	a rash	style of this	suffocated	towels to	to use towels
gash	edges of	description into your	bled	the cleaner	many times these	valances	color height	back a little
blade	cut your	edges of scissors	inflamed	longer bleed	reaction to this	giraffes	height weight	buckle sometimes difficult
finger	cut myself	that dig cut	fairness	leaks or	itchy from being	fitted	pull car	good for breathing
cuts	edges to	about the edges	armour	white splashes	biochemical or cellular	targets	one sided	air passage check
blood	the edges	which needed stitches	autoimmune	rash why	sick to my	stacker	sons chin	wobbly than i

Bruise injury			Burn injury			Poisoning injury		
Word	Phrases		Word	Phrases		Word	Phrases	
	2-word	3-word		2-word	3-word		2-word	3-word
marks	marks on	arms will be	carbon	so hot	a chemical burn	pneumonia	was lead	the hospital along
bruises	red marks	heavy for a	melamine	glass door	got so hot	mold	small plastic	lead poisoned and
bruise	pinched its	of pure fear	agony	chemical burn	like a chemical	edema	blocks stay	the doctors discovered
woods	that dangerous	my already injured	flammable	when browning	a wet cool	poisoned	trash can	mold and trace
scratches	scratches on	went over the	fanning	hamburger it	towel wrapped around	amounts	mega magnetic	designs with small
sores	discover both	and bruising my	sausage	kernels at	by wiping them	powerful	push for	and designs with
handlebars	blade it	has been recalled	steaming	dedicated griddle	fire when the	aspiration	build different	our own home
bruising	had red	the rental vendor	uncooked	a wet	time out was	warns	sheets and	came home and
jamming	it created	facing it left	brownish	fanning the	kernels actually pop	agent	washing we	subsequent aspiration into
deemed	bruises easily	with a bruised	dioxide	motor fan	the kernels actually	trace	designs with	magnets the box

Appendix C. Number of true positives in top-scored and bottom-scored sentences, for each code value of event, hazard, and injury categories

Table 9. Number of true positives for each code value of *event* category in top 200- and bottom 200-ranked sentences, for each scoring method

(a) Amazon

Scoring method	Pathway to injury	Potential injury	Actual injury	Injury mitigation	Product remediation
<i>AFINN</i>					
Bottom 200	21	4	4	5	6
Top 200	31	17	20	1	2
<i>Unigram</i>					
Bottom 200	1	0	0	0	2
Top 200	68**	42**	40**	11**	14**
<i>Bigram</i>					
Bottom 200	18	7	1	0	0
Top 200	71**	27**	37**	11**	16**
<i>Trigram</i>					
Bottom 200	20	7	4	0	0
Top 200	67**	30**	28*	11**	15**

(b) Fortune 50 retailer

Scoring method	Pathway to injury	Potential injury	Actual injury	Injury mitigation	Product remediation
<i>AFINN</i>					
Bottom 200	15	5	2	3	-
Top 200	21	9	6	0	-
<i>Unigram</i>					
Bottom 200	25	4	8	2	-
Top 200	30*	16**	19**	1	-
<i>Bigram</i>					
Bottom 200	22	3	8	2	-
Top 200	24	15*	8	0	-
<i>Trigram</i>					
Bottom 200	22	3	8	2	-

Top 200	22	12	7	0	-
---------	----	----	---	---	---

*Smoke scoring methods (n-grams) significantly differ from sentiment analysis ($p < .05$)

** Smoke scoring methods (n-grams) significantly differ from sentiment analysis ($p < .01$)

- None

Table 10. Number of true positives for each code value of *hazard* category in top 200- and bottom 200-ranked sentences, for each scoring method.

(a) Amazon

Scoring method	Choking	Chemical	Fire	Fall	Hitting	Mechanical	Strangulation	Suffocation	Swallowing	Thermal
<i>AFINN</i>										
Bottom 200	2	1	1	9	4	2	0	1	1	0
Top 200	18	9	2	12	5	3	4	1	1	0
<i>Unigram</i>										
Bottom 200	0	0	0	1	0	0	0	0	0	0
Top 200	28**	19**	11**	42**	20**	12**	10**	8**	5**	2
<i>Bigram</i>										
Bottom 200	1	3	0	2	0	1	0	0	0	0
Top 200	26*	9	11**	43**	19**	7*	10**	9**	5**	2
<i>Trigram</i>										
Bottom 200	0	1	0	4	0	0	2	0	0	0
Top 200	28**	17**	11**	46**	20**	12**	2	9**	5**	0

(b) Fortune 50 retailer

Scoring method	Choking	Chemical	Fire	Fall	Hitting	Mechanical	Strangulation	Suffocation	Swallowing	Thermal
<i>AFINN</i>										
Bottom 200	0	0	3	1	6	5	-	-	0	4
Top 200	1	32	25	0	0	6	-	-	0	1
<i>Unigram</i>										
Bottom 200	0	0	4	0	2	0	-	-	0	0
Top 200	2**	30	90**	13	32	48**	-	-	3	30**
<i>Bigram</i>										
Bottom 200	0	1	10	0	1	1	-	-	0	0
Top 200	2**	33	67**	13	30	47**	-	-	3	29**
<i>Trigram</i>										

Bottom 200	0	0	6	0	2	0	-	-	0	0
Top 200	2**	32	81**	9	27	45**	-	-	3	29**

*Smoke scoring methods (n-grams) significantly differ from sentiment analysis ($p < .05$)

** Smoke scoring methods (n-grams) significantly differ from sentiment analysis ($p < .01$)

- None

Table 11. Number of true positives for each code value of *injury* category in top 200- and bottom 200-ranked sentences, for each scoring method

(a) Amazon

Scoring method	Allergy	Asphyxia	Bruise	Burn	Cut	Poisoning
<i>AFINN</i>						
Bottom 200	0	1	1	-	1	0
Top 200	3	3	4	-	4	2
<i>Unigram</i>						
Bottom 200	0	0	0	-	0	0
Top 200	3	7*	7	-	12**	3
<i>Bigram</i>						
Bottom 200	0	1	1	-	1	0
Top 200	2	5	4	-	8*	4
<i>Trigram</i>						
Bottom 200	0	1	1	-	4	0
Top 200	1	3	2	-	3	4

(b) Fortune 50 retailer

Scoring method	Allergy	Asphyxia	Bruise	Burn	Cut	Poisoning
<i>AFINN</i>						
Bottom 200	-	-	-	2	1	1
Top 200	-	-	-	1	2	0
<i>Unigram</i>						
Bottom 200	-	-	-	0	0	0
Top 200	-	-	-	17**	22**	0
<i>Bigram</i>						
Bottom 200	-	-	-	0	1	0
Top 200	-	-	-	17**	17**	2

<i>Trigram</i>						
Bottom 200	-	-	-	0	3	0
Top 200	-	-	-	17**	4	4

* Smoke scoring methods (n-grams) significantly differ from sentiment analysis ($p < .05$)

** Smoke scoring methods (n-grams) significantly differ from sentiment analysis ($p < .01$)

- None

Appendix D. Comparison of the smoke scoring methods with the conventional sentiment method, AFINN, in detecting the code values across the top N-ranked sentences for each code value in event, hazard, and injury categories.

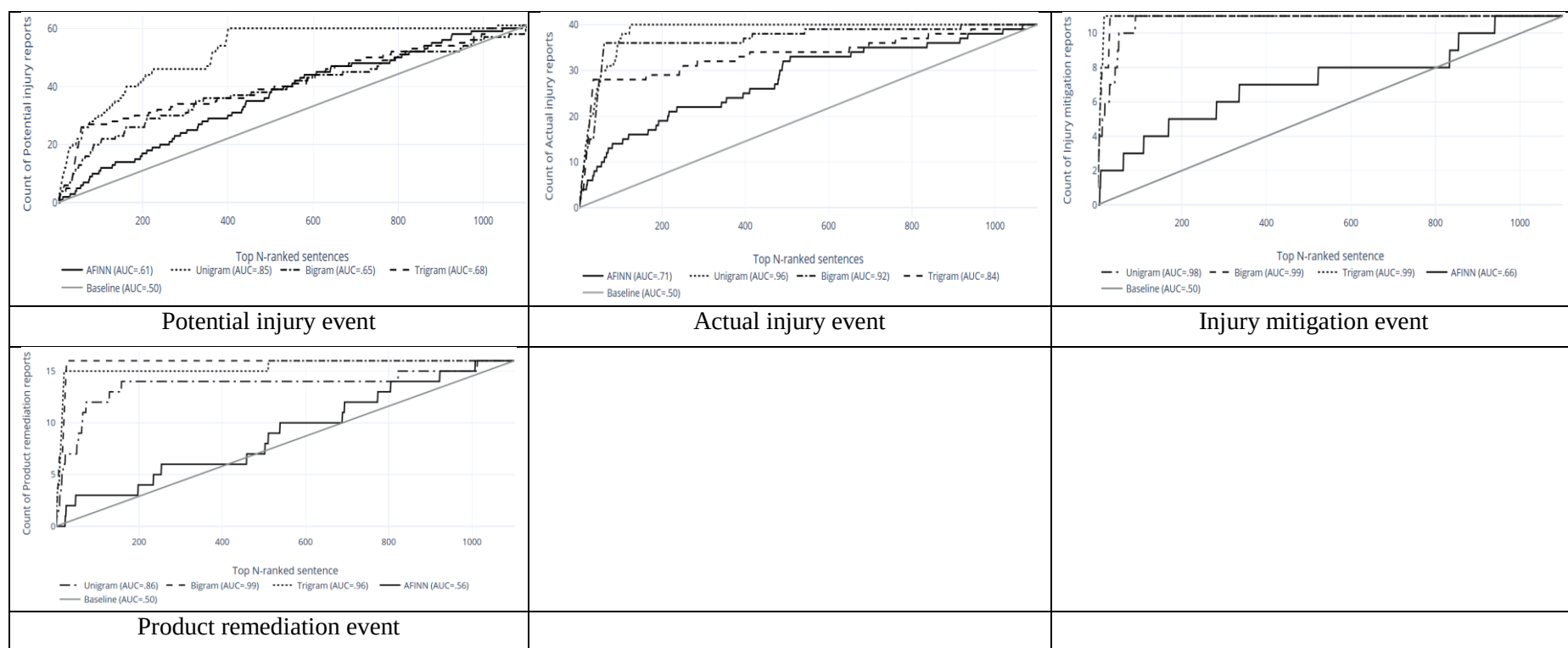


Figure 12. ROC curves for each code value of event category, comparing smoke scoring methods and sentiment approach (AFINN) on Amazon data set.

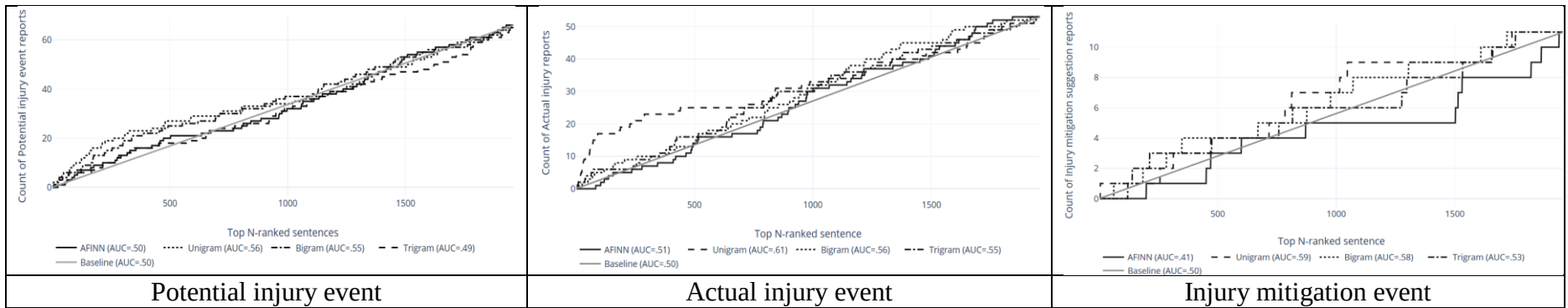
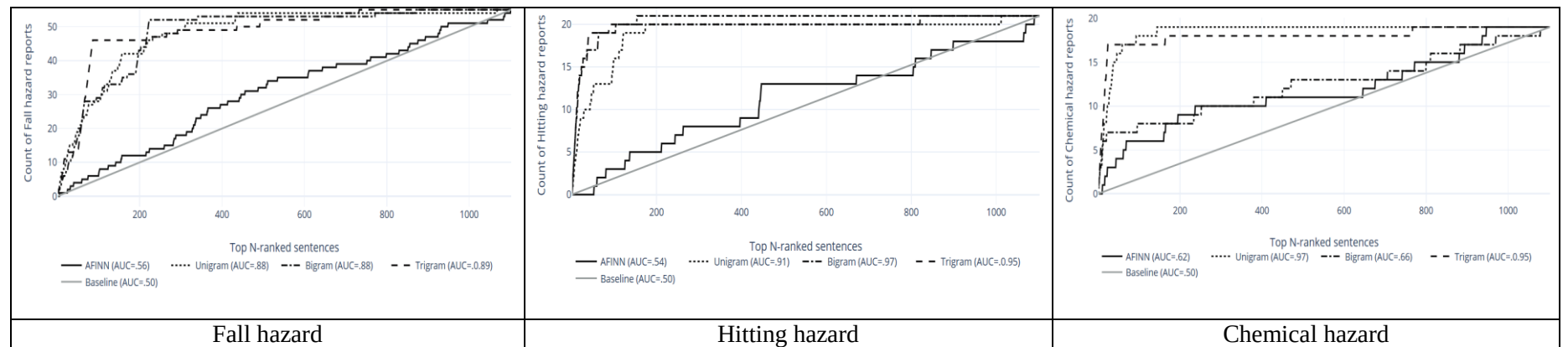


Figure 13. ROC curves for each code value of event category, comparing smoke scoring methods and sentiment approach (AFINN) on the Fortune 50 retailer data set.



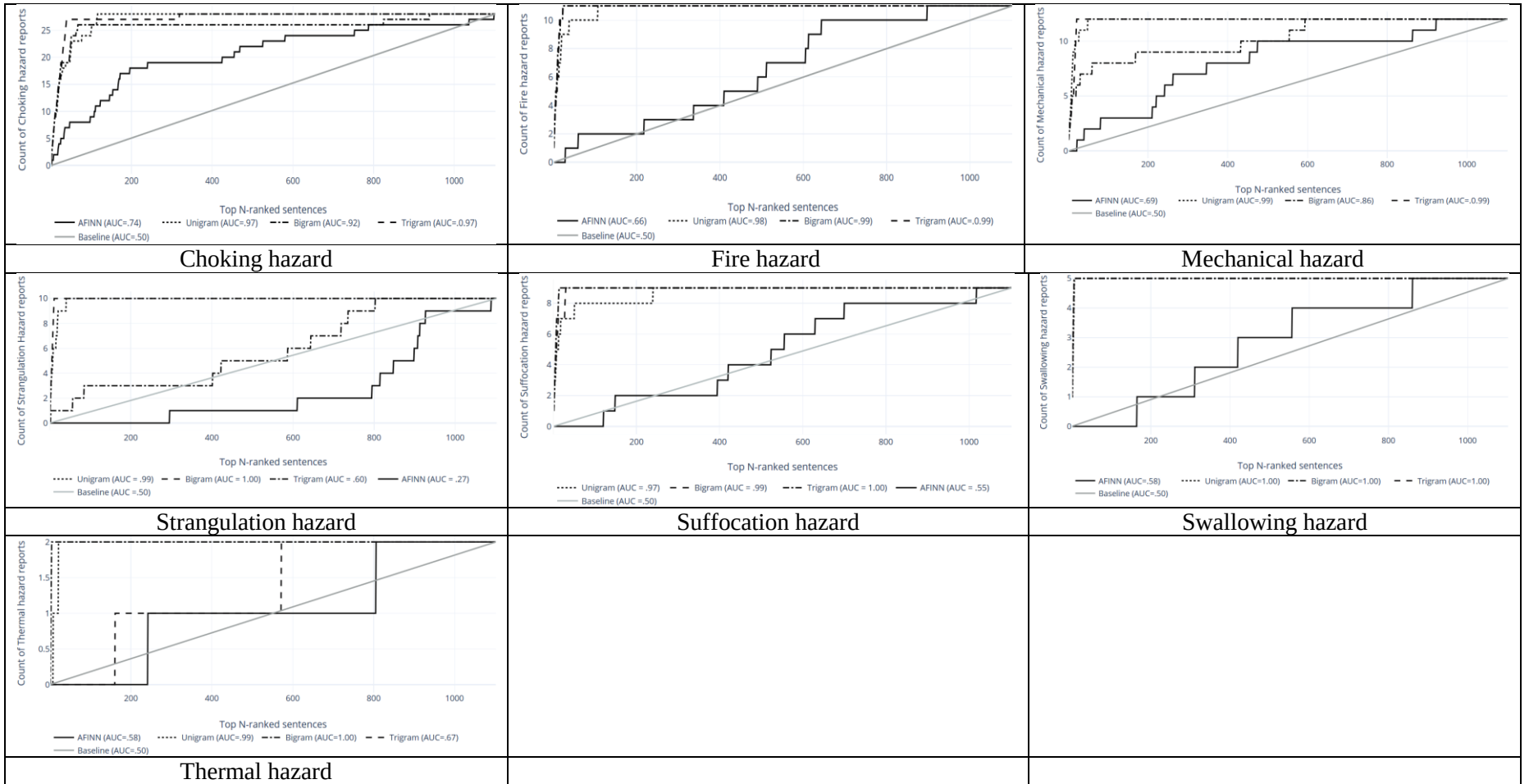


Figure 14. ROC curves for each code value of *hazard* category, comparing smoke scoring methods and sentiment approach (AFINN) on the Amazon data set.

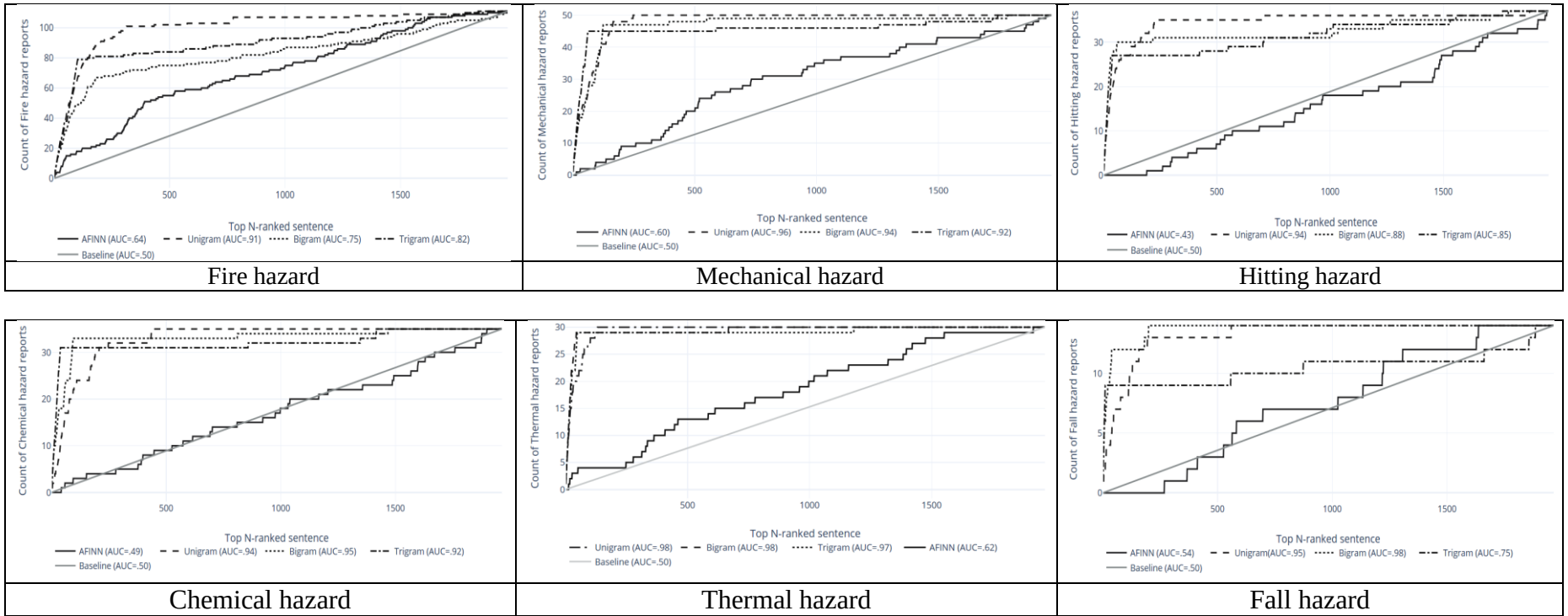


Figure 15. ROC curves for code values of *hazard* category, comparing the performance of smoke scoring methods and sentiment approach (AFINN) on the Fortune 50 retailer data set.

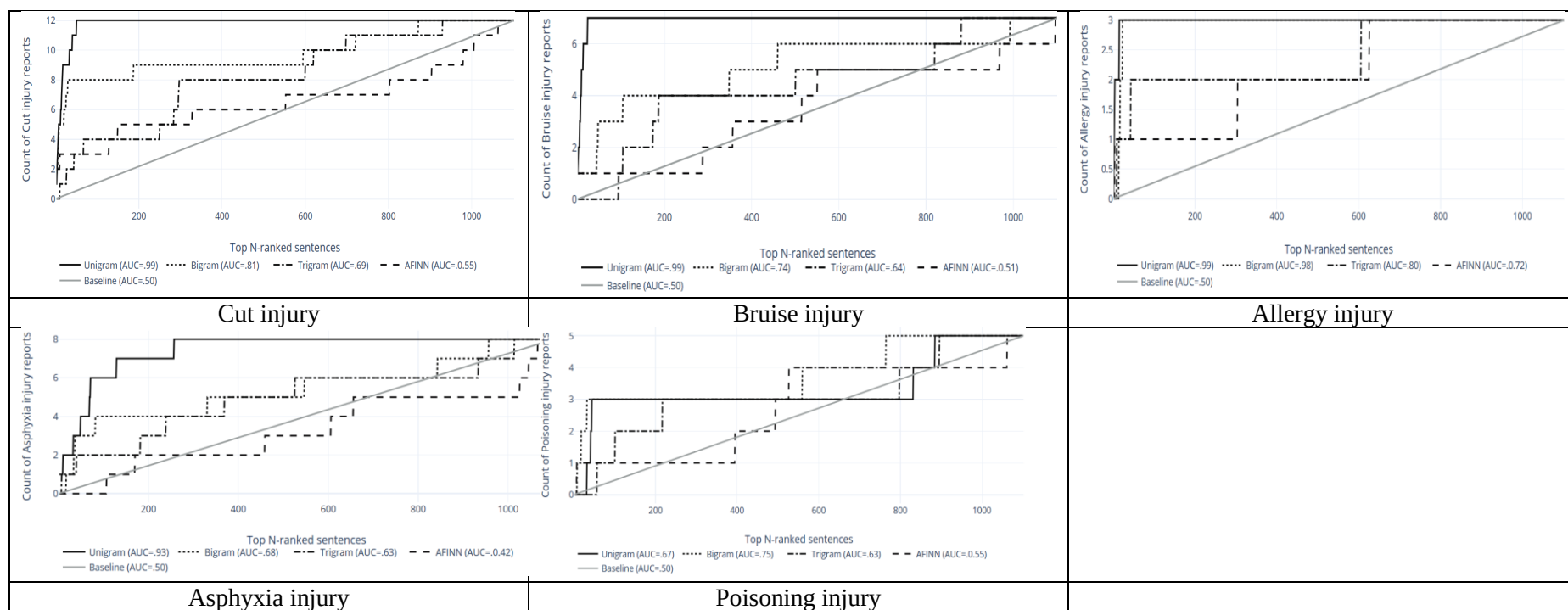


Figure 16. ROC curves for code values of *injury*, comparing the performance of smoke scoring methods and sentiment approach (AFINN) on the Amazon holdout set.

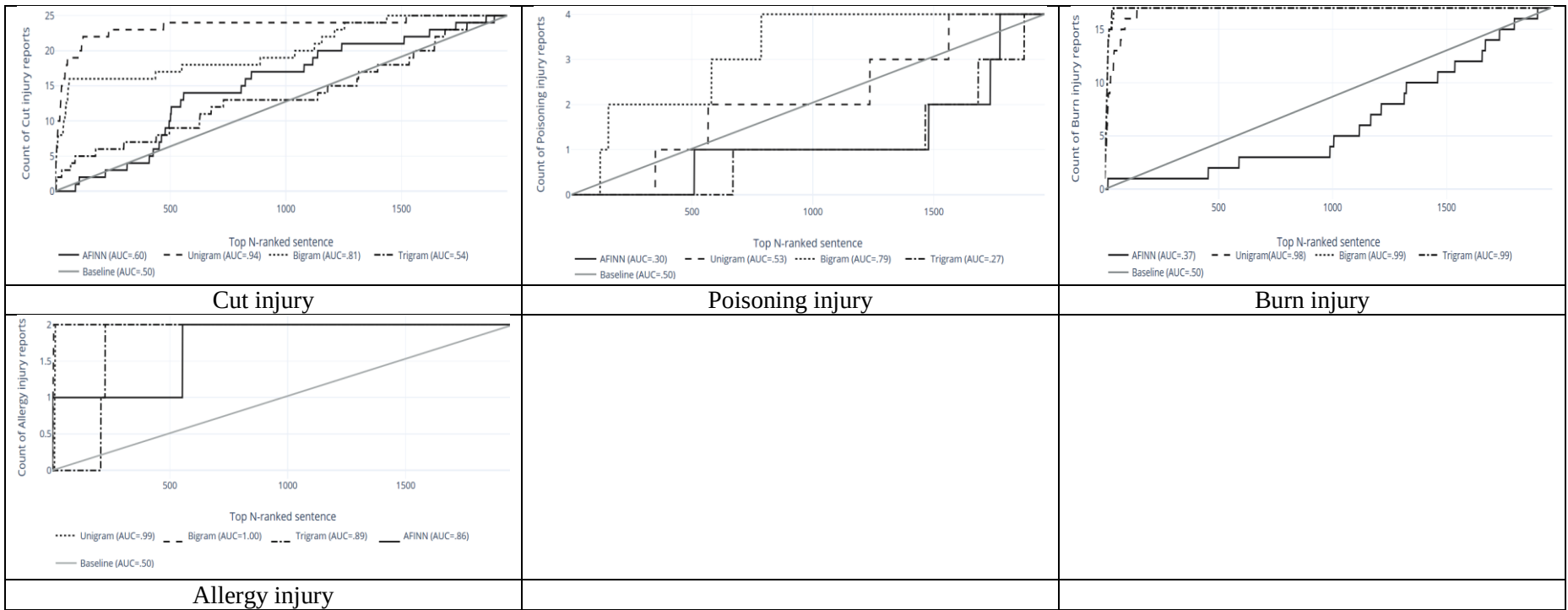


Figure 17. ROC curves for code values of *injury*, comparing the performance of smoke scoring methods and sentiment approach (AFINN) on the Fortune 50 retailer data set.

Appendix E. Assessment of smoke scoring methods in discovering code values of event, hazard, and injury categories across different datasets

In each of the Tables in this Appendix:

- The best performing method on each dataset, for each code value, is indicated in **bold**, if there are sufficient positives (more than 20) to establish dominance with some certainty.
- Parentheses indicate number of available positives for that code value.
- Where the number of available positives for a code value is lower than 20 positives, results are shown **greyed**, to indicate too little data for significance.

Table 12. AUC scores of the smoke scoring techniques in discovering code values of event category across different data sources.

Source	n-gram	Event category					
		Pathway to injury	Potential injury	Actual injury	Injury mitigation	Product remediation	Treatment
Amazon	Unigram	.81 (125)	.85 (60)	.96 (40)	.98 (12)	.86 (16)	- (0)
	Bigram	.71 (125)	.65 (60)	.92 (40)	.99 (12)	.99 (16)	- (0)
	Trigram	.68 (125)	.68 (60)	.84 (40)	.99 (12)	.96 (16)	- (0)
Fortune 50 retailer	Unigram	.60 (152)	.56 (60)	.61 (50)	.59 (10)	- (0)	- (0)
	Bigram	.53 (152)	.55 (60)	.56 (50)	.58 (10)	- (0)	- (0)
	Trigram	.51 (152)	.49 (60)	.55 (50)	.53 (10)	- (0)	- (0)

Table 13. AUC scores of the smoke scoring techniques in discovering code values of hazard category across different data sources

Source	n-gram	Hazard category									
		Fire	Strangulation	Mechanical	Hitting	Chemical	Thermal	Fall	Swallowing	Choking	Suffocation
Amazon	Unigram	.98 (12)	.99 (10)	.99 (12)	.91 (20)	.97 (18)	.99 (2)	.88 (55)	1.00 (5)	.97 (28)	.97 (10)
	Bigram	.99 (12)	1.00 (10)	.86 (12)	.97 (20)	.66 (18)	1.00 (2)	.88 (55)	1.00 (5)	.92 (28)	.99 (10)
	Trigram	.99 (12)	.60 (10)	.99 (12)	.95 (20)	.95 (18)	.67 (2)	.89 (55)	1.00 (5)	.97 (28)	1.00 (10)
Fortune 50 retailer	Unigram	.91 (100)	- (0)	.96 (50)	.94 (40)	.94 (35)	.98 (30)	.90 (15)	.97 (3)	1.00 (2)	- (0)
	Bigram	.75 (100)	- (0)	.94 (50)	.88 (40)	.95 (35)	.97 (30)	.98 (15)	.95 (3)	1.00 (2)	- (0)
	Trigram	.82 (100)	- (0)	.92 (50)	.85 (40)	.92 (35)	.77 (30)	.75 (15)	.97 (3)	1.00 (2)	- (0)

(): value in brackets is the number of reports of the target class available in the dataset (i.e. number of positives, for that code value, in the dataset)

(): value in brackets with black font indicates AUC curve is robust where total number of reports for the target class (i.e. positives) is greater than 20

(): value in brackets with light grey font indicates AUC curve is more brittle (less reliable) because total number of reports in the target class is less than 20 (i.e. total positives, for that code value, in the dataset, is less than 20)

- (0): indicates no AUC available due to no reports of that target class in that dataset (i.e. no positives, for that code value, in the dataset)

Table 14. AUC scores of the smoke scoring techniques in discovering code values of *injury* category across different data sources

Source	n-gram	Injury category					
		Cut	Bruise	Allergy	Asphyxia	Burn	Poisoning
Amazon	Unigram	.99 (12)	.99 (8)	.99 (3)	.93 (8)	.97 (1)	.67 (5)
	Bigram	.81 (12)	.74 (8)	.98 (3)	.68 (8)	.38 (1)	.75 (5)
	Trigram	.55 (12)	.64 (8)	.80 (3)	.63 (8)	.37 (1)	.63 (5)
Fortune 50 retailer	Unigram	.94 (25)	- (0)	.99 (2)	.92 (2)	.98 (17)	.52 (4)
	Bigram	.81 (25)	- (0)	1.00 (2)	.61 (2)	.99 (17)	.79 (4)
	Trigram	.54 (25)	- (0)	.86 (2)	.83 (2)	.99 (17)	.30 (4)

(): value in brackets is the number of reports of the target class available in the dataset

(): value in brackets with black font indicates AUC curve is robust where total number of reports for the target class is greater than 20

(): value in brackets with light grey font indicates AUC curve is more brittle (less reliable) because total number of reports in the target class is less than 20

- (0): indicates no AUC available due to no reports of that target class in that dataset

Chapter 2. A Semi-automated Risk Assessment Method for Consumer Products

Abstract

In this study, we develop a model that assesses product risk using online reviews from Amazon.com. Beyond using smoke terms to identify hazards, we estimate risk severity using hazard-type weights, and risk likelihood using total reviews as a proxy for sales volume. In addition, we obtain expert assessments of product hazard risk (risk likelihood and severity) for a sample of high- and low-risk consumer products identified by a computerized risk assessment model we have developed. The experts' responses are utilized for calibrating and validating our computerized product risk-assessment scoring model. This model helps businesses determine the risk severity, risk likelihood, and overall risk level of a specific product. The model produced by this study could be helpful for product safety practitioners in product risk identification, characterization, and mitigation.

1. Introduction

Regulators and consumers are concerned about potential product safety issues. Safety concerns and incidents may lead to a product recall. Recalls may cost business millions of dollars.

Consumers may receive medical treatment in emergency departments¹¹ because of “consumer product-related injuries”. For instance, in Australia, the cost is estimated to range between \$194 and \$238 million annually which includes the direct treatment of consumer product-caused

¹¹ Bruce A. Lawrence, Ted R. Miller, Geetha M. Waehrer, Rebecca S. Spicer, Mark A. Cohen, William W. Zamula, The Consumer Product Safety Commissions Revised Injury Cost Model, February 2018. Available from: <https://cpsc.gov/s3fs-public/ICM-2018-Documentation.pdf?YWuW4Jn0eb2hExeA0z68B64cv6LlUYoE>

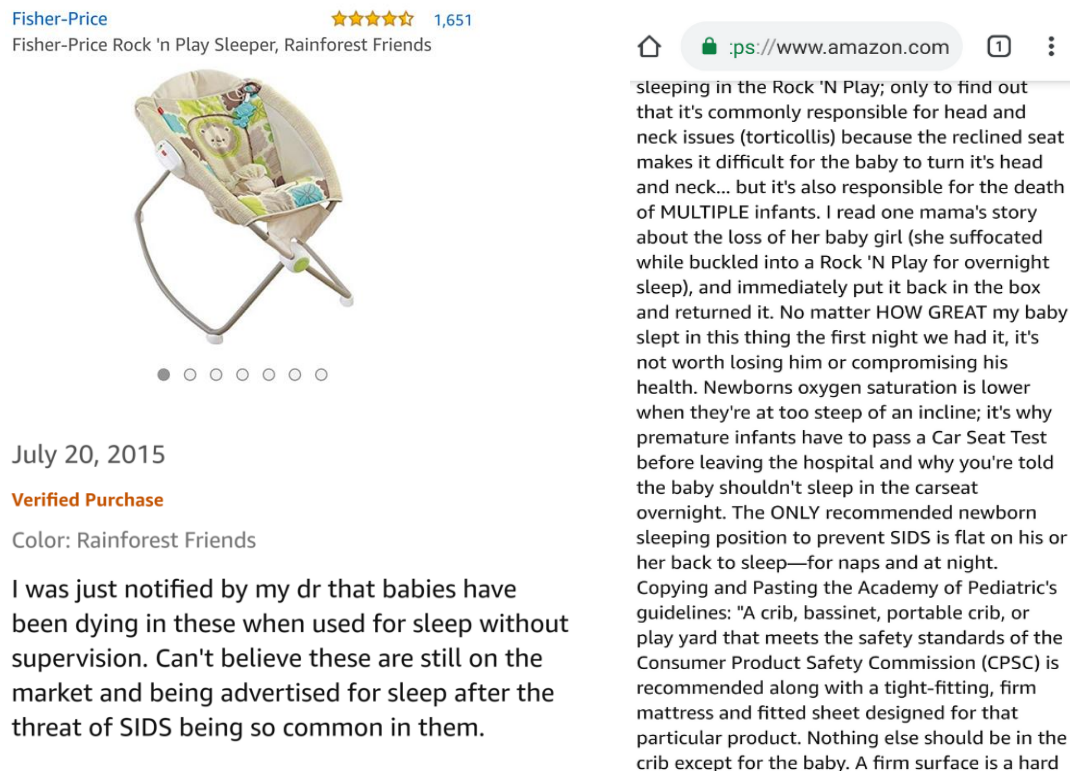
injury (Watson and Ozanne-Smith, 1995). This estimate implies there is a high expense associated with medical and hospital treatment of consumer product-related injuries.

According to InjuryFacts.nsc.org, 10.5 million people were treated in emergency departments in 2017, due to consumer product-related injuries. Consumer product-related injuries increased by 4%, i.e. from 10,066,530 in 2016 to 10,507,009 in 2017. Additionally, injuries increased by 9% in 2014. Most of these injuries fall under two specific segments of the population: older adults and young children. Commonly used consumer products in the market such as televisions, household cleaning products and also beds pose risks to various hazards and injuries. Children aged from 0 to 4, have the highest injury rates of any age group¹².

Recently, for example, Consumer Reports (CR) investigated safety concerns related to the Fisher-Price Rock 'n Play Sleeper (Peachman, 2019). This baby product was made for babies to sleep in an inclined position. The investigation revealed that there were at least 32 infant deaths since 2009. Even though there were several deaths, and safety concern of this specific product, the CPSC and Fisher-Price issued the first public warning about the product only on April 5th 2019 (CPSC, 2019a), and the company did not recall the product from the market until April 12, 2019, a number of years after the initial death reports (CPSC, 2019b). Interestingly, as early as 2015, customer had posted online reviews of this product on Amazon, mentioning baby deaths – see Figure1 below.

¹² Consumer Product Injuries. (2017). Retrieved from <https://injuryfacts.nsc.org/home-and-community/safety-topics/consumer-product-injuries/>

Figure 1: Amazon reviews for Fisher Price Rock 'n Play Sleeper from July 20, 2015, and after, mentioning baby deaths – both postings prior to CPSC warning and recall in April 2019



The CPSC is unable to monitor every product on the market and usually takes actions on consumer products with safety concerns after multiple hazardous incidents and injury events have already occurred (Nasri et al., 2018). Therefore, CPSC should take steps in obtaining and using reliable data for rapidly detecting potential hazards and assessing product safety-related risks. Additionally, CPSC should undertake an automated approach focused on reducing the severity of an injury and should use these automated techniques to assist with making faster and more targeted recommendations in the area of consumer product safety. As online reviews are an open-source platform to the public prior to online purchasing, regulators and consumers should take advantage of this resource to rapidly assess different risks associated with specific products.

In December 2016, a Pew Research Center survey¹³ found that 79 percent of U.S. adults purchased a product directly through a social media platform. When the Center first surveyed online shopping in June 2000, only 22 percent of Americans purchased through online media. Hence, there has been significant growth of online shopping in the U.S. Additionally, customers use online social media (e.g. Facebook), product reviews on retailer websites (e.g. Amazon.com, Walmart.com), and discussion forums (e.g., Patient.info, Toyotanation.com) for sharing their complaints with other customers regarding product defects (Abrahams et al., 2012; Zhang et al., 2019) and safety hazards (Adams et al., 2017; Nasri et al., 2018). Other online data sources relating to product safety hazards also exist. For instance, the CPSC generates annual National Electronic Injury Surveillance System (NEISS) data that summarize injury data according to major product groups, age, and sex. The 2017 NEISS dataset¹⁴ currently stores over 380,000 brief narratives describing people's incident scenarios. These narratives contain limited content related to a particular hazard discussion of a consumer product. In contrast, due to their rich content, there is much flexibility when it comes to using online data sources, such as Amazon and Facebook, for extracting useful hazard-related information.

¹³Anderson, A. S. a. M. (2016). Online Shopping and E-Commerce. Retrieved from: <https://www.pewinternet.org/2016/12/19/online-shopping-and-e-commerce/>

¹⁴ CPSC. National Electronic Injury Surveillance System (NEISS). Retrieved from: <https://www.cpsc.gov/Research--Statistics/NEISS-Injury-Data>

NEISS	Amazon.com
58 YOM went to bathroom; lost his balance, fell to floor hit upper back on toilet.	Bought the item for my mother. She is little unsteady on her feet and fell off this due to fact <i>not secured to rim. It tips due to the person leaning forward to get off</i> causing them to go off balance...for older people <i>it cannot just slide unit must have secure attachment other than few sticky pads.</i>

Table 1: Examples of Hazard Reports (**Bold:** Hazard Type; *Italic:* Product Safety Concerns) from Two Different Data Sources

Table 1 shows two hazard reports collected from two different data sources, namely, the NEISS and Amazon datasets. The NEISS data contain a short narrative mentioning the specific hazard type (in bold), and some basic context, such as the victim's age (58), gender (M), and brief details of the circumstance in which the injury occurred (“went to bathroom; lost his balance”). In contrast, the Amazon review contains not only the hazard type but also may contain some additional, free form consumer-written narrative such as the hazard circumstances (“mother ... little unsteady”), product defect description (“not secured to rim...tips due to person leaning... cannot just slide”), and remediation suggestions (“must have secure attachment other than a few sticky pads”). These portions, in italics, illustrate the specific safety concerns associated with a hazard, for example, a fall hazard. Hence, introducing a text analytic approach in our risk-assessment model would benefit safety regulators and product designers for extracting these sentences (shown in italics) and estimating the risk level of the product.

Prior studies have mentioned that many complaints are voiced in online reviews, which are venues for customer feedback, and identification of safety hazards (Mummalaneni, et al. 2018; Nasri et al., 2018). Thus, detecting and monitoring different types of hazards and assessing product risks in online reviews has vital implications for stakeholders like safety regulators and

product designers, as well as professional consumer review organizations, such as Consumer Reports.

The CPSC developed a strategic plan for accomplishing a set of goals and objectives in identifying hazards. We have enlisted two CPSC goals, namely, “rigorous hazards identification” and “decisive response” (see Table 2).

Goal 1: Rigorous hazards identification	Goal 2: Decisive response
1.1. Reduce the time of identifying hazard trends by improving the assessment of hazard-related data	2.1. Extend the CPSC’s functionality to inspect and monitor defective products
1.2. Develop a risk-assessment methodology to detect and prioritize hazards	2.2. Utilize a risk-based method to prioritize the CPSC’s response to product hazards
1.3. Monitor the market regularly to track whether previously identified hazards exist in similar products	2.3. Increase the speed of recalls of noncompliant and defective products
1.4. Increase hazard surveillance of consumer products to discover recalled products	2.4. Decrease the time of informing customers about newly detected hazards and appropriate actions to take

Table 2: CPSC Strategic Plan¹⁵

This study focuses on accomplishing goals 1.1-1.3 under the rigorous hazards identification category, as well as goal 2.1-2.4, under the category of decisive response. Chapter 1 untangled the various types of narratives related to safety concerns regarding hazard types across diverse product categories, and this study (Chapter 2) now uses the finely labeled text to estimate an overall risk level for the product. To achieve the goals prescribed by the strategic plan of CPSC (see Table 2 in bold), we adopt a text-classification method for revealing hazard-specific reports from online (real-time) reviews. We then incorporate this method into a risk-assessment scoring model. This model assesses product risk using online reviews from Amazon.com. The risk scores generated by our model are compared with scores assigned by the product safety experts

¹⁵ CPSC. (2016). *Strategic Plan*. Retrieved from: https://www.cpsc.gov/s3fs-public/2011strategic_1.pdf

completing an online survey (11 products assessed per survey). This model also helps stakeholders prioritize risk by computing risk scores and risk levels of a certain product. In addition, this study aims to conduct expert assessments of product hazard risk (hazard likelihood and severity) for a sample of high- and low-risk consumer products identified by the computerized risk-assessment model we developed. The experts' responses are utilized for calibrating and validating the computerized product risk-assessment scoring model.

2. Literature Review

This section explains related work on using structured data sources, online reviews, and text classification.

2.1. Structured data

Many developed countries have organized risk surveillance, risk assessment, and risk alerting systems for monitoring and estimating product risk in light of reported safety incidents related to consumer products, and informing stakeholders of these risks. Some examples of such systems are NEISS by the CPSC (Paston, 2018) in the United States, Correction Action Plan (CAP)¹⁶ by the British Standards Institution (BSI) in the United Kingdom, Risk Characterization Methodology (RCM)¹⁷ by the Health Canada in Canada, Australian/NZ Nomograph by the Australian Competition Consumer Commission (ACCC)¹⁸ in Australia,

¹⁶ BSI. *Corrective action: The closed-loop system*. Retrieved from: <https://www.bsigroup.com/LocalFiles/en-GB/entropy/BSI-Corrective-and%20Preventive-Actions-Whitepaper-EN-GB-UK.pdf>

¹⁷ Health Canada. *Health Canada Decision-Making Framework for Identifying, Assessing, and Managing Health Risks*. Retrieved from: <https://www.canada.ca/en/health-canada/corporate/about-health-canada/reports-publications/health-products-food-branch/health-canada-decision-making-framework-identifying-assessing-managing-health-risks.html>

¹⁸ A. C. C. C. *ACCC Product Safety Nomograph Instructions for Use*. Retrieved from: <https://foi.accc.gov.au/sites/foi.accc.gov.au/files/repository/PS%20%20ACCC%20Product%20Safety%20Nomograph%20Tool%20and%20Instructions%20for%20Use%20%284%29.doc>

Rapid Information System (RAPEX)¹⁹, Risk Assessment Model (RAM) (Union, 2016), and Risk Assessment Guidelines (RAG)²⁰ in the European Union. Table 3 shows that each system falls under the following functionalities:

Risk Assessment	Risk alerting	Risk surveillance
- ACC Nomograph; - BSI CAP; - Health Canada RCM; - EU RAG; - EU RAM;	- EU RAPEX - CPSC Recalls List (cpsc.gov/recalls)	- CPSC NEISS

Table 3. Functionalities of each safety regulatory agency

These national product safety agencies effectively store information on safety hazard–related incidents. Furthermore, the data on a large number of these incidents are useful for analyzing hazard risk. Following are some illustrative examples of structured data sources utilized by government agencies for identifying the product safety risks, and informing decisions about recalls, policy-making, or updating technical standards.

2.1.1. Consumer safety complaints dataset²¹ (Saferproducts.gov)

Saferproducts.gov is a national reporting system maintained by the CPSC that allows consumers to report safety-related product concerns to the organization. The Saferproducts.gov dataset includes attributes like the incident narrative/description, location, product categories, manufacturer name, age, gender, injury reports of consumers and treatment events of an injury. The archived data source records consumer complaints,

¹⁹ Union, E. Guidelines for the management of the European Union Rapid Information System ‘RAPEX’. Retrieved from: https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/repository/content/pages/rapex/docs/Guidelines%20annex_en.pdf

²⁰ Union, E. Risk Assessment for RAPEX. Retrieved from: <https://ec.europa.eu/consumers/consumer-safety/rag/>

²¹ <https://www.saferproducts.gov/Search/default.aspx>

for example, a complaint reported on April 27, 2018: “[W]hen I set it down it went off and took finger nails and tips of 2 fingers off.”

2.1.2. Hospitalization dataset²² (NEISS)

NEISS is a publicly accessible dataset maintained by the CPSC covering a national sample of product-related hospitalization incidents. This dataset includes the following attributes: incident narrative, treatment date, age, gender, body part, diagnosis code, location, and degree of fire involvement. Each narrative only mentions a specific event of an injury. The dataset stores injury reports of patients who received treatment in hospitals, for example, “71YOM FELL ON FLOOR IN BATHROOM. DX²³ RIB CONTUSION.”

2.1.3. Product utilization dataset²⁴ (Nielsen)

Nielsen maintains a dataset detailing the utilization of products by consumers across the United States. This dataset includes attributes like sales volume, product descriptions, and market penetration. This structured data source is constructed from consumer surveys administered to stratified random samples. The dataset contains around 120,000 households that geographically represent the most recent US Census. Though not directly related to product safety, the Nielsen dataset could be helpful as a convenient 3rd

²² CPSC. National Electronic Injury Surveillance System (NEISS). Retrieved from: <https://www.cpsc.gov/cgibin/NEISSQuery/home.aspx>

²³ NEISS dataset includes many shorthand codes used by the medical coding professionals, e.g. DX is shorthand for “Diagnosis”.

²⁴ Panel, N. C. National Consumer Panel A Nielsen/IRI Joint Venture. Retrieved from: <http://www.ncppanel.com/content/ncp/ncphome.html>

party source for estimating the penetration of popular products in the marketplace, which may inform the risk likelihood of the product in the case where the product presents a safety hazard.

2.1.4. GlobalRecalls portal²⁵

The GlobalRecalls portal stores electronic data about product recalls²⁵. Regulatory agencies globally issue the consumer product recall on the electronic portal and make the information publicly available on a regular basis. Consumers and companies may use this portal to verify the safety concerns of the products that they plan to buy in the future. This portal is also useful when purchasing products online cross borders from various countries such as Australia, Canada, United States, etc. As at April 2019, the GlobalRecalls portal listed over 22,000 recalls. The recalls issued by the regulatory agencies provide incidents based on safety concerns of food products, vehicles, pharmaceuticals and drugs, and so forth.

The attributes that are present in consumer complaints and hospitalization data sources are updated annually or bi-annually. In addition, it is a time-consuming and expensive task for government employees to gather narratives mentioning hazard-related incidents from different hospitals around the country. Furthermore, to protect manufacturer confidentiality during investigations (pursuant to Section 6(b)(5) of the United States' Consumer Product Safety Act; 15 U.S.C. §§ 2051–2089), the public NEISS dataset does not mention the manufacturer name, or the specific product name that caused the injury. Indeed, Congressional testimonies

²⁵ OECD. Global portal on product recalls. Retrieved from: <https://globalrecalls.oecd.org/>

by various stakeholders at the recent hearing entitled, “Protecting Americans from Dangerous Products: Is the Consumer Product Safety Commission Fulfilling Its Mission?” organized by The Consumer Protection and Commerce Subcommittee of the House Committee on Energy and Commerce, have advocated for Section 6(b) to be rescinded, so that consumers may be informed more rapidly by the CPSC of safety concerns presented by specific products (Gregg, 2019). The present study introduces an alternative mechanism for using online reviews to rapidly detect hazard category–specific reports for specific products and adopts text-classification methods to automate the risk assessment of consumer products.

2.2. Online reviews

Unlike structured data sources, online reviews are unstructured data, meaning that the content mentions a mix of positive, neutral, and negative comments on different products. These comments multiply over time when the customers interact with each other by leaving their feedback and complaints on online media. As a result, the volume of online reviews has grown exponentially over the past few decades. Hence, it is challenging to extract essential or crucial problems of a specific product.

Past studies have focused on disentangling online reviews to reveal semantic trends (Abrahams et al., 2012; Goldberg and Abrahams et al., 2018). This work showed that general defects do not usually elicit strong emotions, and therefore, they are not emphasized, even if they are implicit and potentially hazardous. If reviewers want to post unbiased criticism, the wording of the defect reports may be less polarizing than that in other online reviews (Mummalaneni et al., 2018). Similarly, the text of the online reviews may not always be

emotive when it comes to explaining safety hazards, and hence, it becomes difficult for traditional text analysis, such as sentiment analysis for detecting safety hazards (Abrahams et al., 2012, 2013, 2015). For example, online consumers repeatedly post their negative comments on product usage, but most of these negative comments are innocuous complaints about the product (e.g. color, size, value, effectiveness, etc.), and few of these negative comments are concerns regarding safety hazards associated with the product.

2.3. Text classification

A series of past studies concentrated on applying text classification methods to defect discovery and safety surveillance (Abrahams et al., 2015; Abrahams et al., 2012; Adams et al., 2017; Goldberg and Abrahams, 2018; Law et al., 2017; Mummalaneni et al., 2018; Nasri et al., 2018; Winkler et al., 2016). Nasri et al. (2018) investigated online videos of customers who used multiple product categories to detect the existence of a safety hazard in general.

Rather than only identifying narratives related to safety hazards across multiple products, the current study also aims to classify reports indicating various subcategories of safety hazards, such as chemical hazards, choking, falls, and fire. In addition, natural language processing (NLP) rules are utilized in the current study to classify these hazard-specific narratives into subcategories by adapting the n-gram generation techniques from earlier studies (Abrahams et al., 2012, 2013, 2015). Before generating the most important keywords for each subcategory, a human coding (i.e., tagging) procedure for classifying these subcategories is applied to the current study. Instead of classifying only short narratives (Marucci-Wellman et al., 2017), the current study aims to classify a large pool of safety

hazard reports (Vallmuur et al., 2016) into subcategories of safety hazards among multiple products. This is done using the semiautomatic classification approach, that is, human tagging, to classify different categories of hazard reports. Consequently, the current study adopts the semi-automatic n-gram scoring technique (Goldberg and Abrahams, 2018; Nasri et al., 2018) to construct the risk-assessment scoring model for determining the risk levels of different products based on their category-specific hazard reports.

3. Research Contribution and Gap

Table 4 includes past research focused on risk assessment of consumer products.

Study	Hazard type existence	Domain	Risk framework	Risk factor identification	Data source
Pan et al. (2014)	Multiple	Multiple consumer products	x	x	Government agency websites
Wu et al. (2017)	Physical	Toy	x	x	Survey questions
Lee et al. (2018)	Chemical	Household deodorizing product		x	Screening Information Dataset (SIDS) report
Chiarini-Tremblay et al. (2009)	Fall	Hospital		x	Electronic medical record (text)
Lee et al. (2012)	Chemical	Humidifier		x	Survey questions
Bondy et al. (2005)	Fall, Mechanical	Construction workplace		x	Accident investigation reports
Chen et al. (2015)	Fall, Mechanical	Emergency department		x	Medical text data
Our study	Multiple	Multiple consumer products	x	x	Online customer reviews

Table 4: Research Studies on Risk Assessment for Consumer Products

The researchers (see Table 4) used conventional and structured datasets that do not update frequently. The researchers also retrieved risk factors and built the risk assessment framework based on the conventional data sources only. Conversely, online (real-time) reviews grow rapidly in number every day (Zhang et al., 2019). Most of the research studies focused on only one domain to identify risk factors; the one exception is Pan et al. (2014), who investigated various consumer products using the government agency's hazard surveillance data sources, such as EU RAPEX and CPSC NEISS. These datasets also represent conventional data, which are updated with incidents related to different hazard and injury cases once or twice a year.

Additionally, Section 6(b) of Consumer Product Safety Improvement Act (CPSIA) prevents CPSC from disclosing manufacturer's name or model identifier for investigations, which inhibits rapid safety alerting to consumers. Our approach provides an alternatively timely alerting procedure, given the restrictions placed on CPSC by Section 6(b).

Table 5 lists various risk metrics that are used in different practitioners' risk-assessment frameworks. The current study adopts the risk metrics of the risk-assessment flowchart from the EU guidelines for classifying risk level, severity level, and the likelihood of injury. As shown in Table 5, our study adopts the risk metrics (in bold) of the practitioners' risk frameworks to construct our risk-assessment framework for safety hazard surveillance in online reviews. The risk metrics (in italics in Table 5) are new elements that are only available from the online reviews, not in conventional structured data sources. For example, our framework considers the total count of customer reviews of a specific product from Amazon.com, as a proxy for product popularity, and consequently risk likelihood. As Liu, Feng, and Liao (2017) mentioned, the number of online reviews is directly proportional to the sales volume of a product. Hence, we added the review count as one of the elements in our risk-assessment scoring model for determining the overall risk likelihood level of a specific product.

Many empirical studies focus on the effect of online reviews on sales (Chevalier and Mayzlin, 2006; Dellarocas et al., 2007; Duan et al., 2008; Chintagunta et al., 2010). Three measures of online reviews were most discussed: valence, variance, and sales volume (Dellarocas et al., 2007). For instance, considering multiple mainstream reviews, Dhar and Chang (2009) found that the number of online reviews as an independent variable significantly predicted the number of sales in the music industry. Liu, Feng, and Liao (2017) introduced a model combining both online reviews and sales volume in a framework. Likewise, the count of

customer reviews of each product is blended with hazard-specific reports into our risk assessment framework to determine the risk level.

This study contributes to determining the risk level of a product by incorporating two measures from the risk matrix of EU guidelines—likelihood of a specific hazard code type and severity of the hazard code type. The procedure for computing both of these risk measures is mentioned in the Methodology section. Henceforth, the novel overall product risk score computed in this study is referred to as the Zaman-Abrahams-Goldberg-Gruss (ZAGG) risk score, in recognition of the various contributors to its conceptualization.

Risk metric	Risk-assessment framework				
	CPSC ²⁶	Health Canada ²⁷	EU guideline ²⁸	BSI ²⁹	Our study
Hazard type	x	x	x	x	x
Severity level		x	x	x	x
Likelihood of injury	x	x		x	x
Risk level	x	x	x	x	x
Customer type at risk	x	x	x	x	
Age of customer		x			
Age of product		x			
<i>Count of customer reviews</i>					x
<i>Smoke scores of each hazard-specific report</i>					x
<i>Weight of each hazard type</i>					x

Table 5: Practitioner Risk Assessment Frameworks

²⁶Poston, A. (2018). *Guidance on the Application of Human Factors to Consumer Products*. Retrieved from <https://www.cpsc.gov/s3fs-public/HF-Standard-Practice-Draft-12Feb2018.pdf?CGk4Zs9GAbjCnZ5RXQuSlr2toQ1aLPhJ>

²⁷Health Canada. *Health Canada Decision-Making Framework for Identifying, Assessing, and Managing Health Risks*. Retrieved from: <https://www.canada.ca/en/health-canada/corporate/about-health-canada/reports-publications/health-products-food-branch/health-canada-decision-making-framework-identifying-assessing-managing-health-risks.html>

²⁸Union, E. Guidelines for the management of the European Union Rapid Information System ‘RAPEX’. Retrieved from: https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/repository/content/pages/rapex/docs/Guidelines%20annex_en.pdf

²⁹BSI. (2017). *Consumer product safety related recalls and other corrective actions*. Retrieved from: <https://www.theiabm.org/wp-content/uploads/2017/10/PAS-7100-Review-draft-for-reviewer-download.pdf>

Our model is computer-supported (partially automated) for the rapid discovery of subcategories of safety hazards from real-time (online) reviews. One of the contributions is generating n-grams (unigrams, bigrams, and trigrams) for each hazard type. Consequently, this study contributes to constructing the risk-characterization framework based on hazard-specific reviews. In addition, NLP rules are utilized for classifying the safety narratives into specific types of hazard reports by adapting the n-gram generation techniques from earlier studies (Abrahams et al., 2012, 2013, 2015). Before generating the most important keywords for each hazard type, in the Methodology section, a human coding (i.e., tagging) procedure for classifying safety narratives into different hazard types is discussed. The keywords are also known as the category-specific “smoke” terms (Nasri et al., 2018; Winkler et al., 2016). We apply this smoke-scoring methodology to this study for scoring an unseen (holdout) dataset of hazard-specific reports that indicate their relative frequency (prevalence) for a specific hazard type (e.g., chemical hazard) versus nonspecific hazard type (e.g., no chemical hazard). The previous chapter (Chapter 1) found that hazard-category-specific smoke words can be used to subcategorize safety hazards found in online reviews, even across diverse product categories. To generalize our risk assessment framework, we are considering the smoke terms of each hazard type generated from the hazard-specific reports of different product categories. We focus on executing the following research contributions (RCs) in this study:

- RC 1: Constructing a semi-automated risk-assessment model by adopting the risk metrics from practitioners’ frameworks;
- RC 2: Applying the smoke word dictionary of each hazard code type (from Chapter 1) to compute the ZAGG severity and likelihood of each hazard type for determining the risk level of a specific product; and

- RC 3: Showing the applicability and validity of our risk-assessment model by checking the human (expert) judgment of risk level against the automated judgment of the risk assessment of each product.

4. Methodology

The following subsections describe the data collection, data coding and processing, sentence-level tagging procedure, generation of smoke words, and risk scoring procedure.

4.1. Data collection

The availability of reliable and representative data is a primary requirement for effectively retrieving specific information on hazard-related reports. As a first step, the hazard-related reports were obtained from Amazon.com. The Amazon.com dataset consisting of reviews of 17 product categories (see Chapter 1). We randomly collected the customer reviews of these product categories; as a result, they represent a diverse section of products that customers purchase via the internet. This diverse sample allows us to study the product safety hazard occurrences across multiple industries, spanning from baby products to countertop appliances, health/personal care, sports equipment, and others.

4.2. Data coding and processing

Coding was undertaken for the Amazon.com dataset. Students at a Tier-1 research university manually tagged a random set of reviews by reading and labeling each review in accordance

with whether a safety hazard was mentioned. In total, we studied 124,289 unique reviews (refer to Table 2 of Chapter 1) across 17 product categories. We gave instructions to each undergraduate student to tag a random set of 200 reviews. The undergraduates tagged a total of 181,999 tags (see Table 2 of Chapter 1, Column 2 grand total; some reviews were tagged more than once). 4,938 reviews, or 3.97% of the total, were tagged as referring to a safety concern (refer to Table 2 of Chapter 1, Column 4 grand total).

To verify the tagging accuracy of the students, 4,938 reviews were randomly distributed to a second group of undergraduates, who indicated whether a safety concern incident was mentioned. As explained in detail under the subsection (3.1) of Chapter 1's Methodology section, a second set of undergraduates labelled only 1,389 reviews as safety concern reports out of the 4,938 reviews. Then in a third round, for assessing the tagging accuracy of the undergraduates, graduate students (including the lead author) verified that 740 out of the 1,389 reports truly represented as safety concern narratives. Table 3 of Chapter 1 (Amazon reviews sub-total) shows the distribution of the 740 safety-concern reports among the product categories that were utilized in the analysis of this current study.

In a prior study, Mummalaneni et al. (2018) confirmed that the proportion of online discussions mentioning safety concerns explicitly is quite small. As such, this large pool of data collection is crucial for producing a sample size suitable for studying a set of hazard-specific narratives for a variety of products. Hence, these 740 hazard-tagged reviews were finalized to analyze the risk level of each product in the remainder of this study for the specific categorization of different hazard types. According to the prior study's suggestion (Chapter 1), we focused on sentence-level tagging instead of review-level tagging. This study disentangles the hazard-tagged reviews by manually tagging each sentence of these reviews

for revealing specific discussions of hazard categories across a diverse set of 10 product categories.

4.3. Sentence-level tagging procedure

This study adapted the coding instructions from Health Canada's RCM³⁰ and the EU RAPEX³¹ guideline for manually categorizing different types of hazard. The tagging protocol (taken from Chapter 1) elaborates on the definitions of each hazard type, along with examples extracted directly from online data. To execute this tagging assignment, a group of students conducted tagging of the focal sentences of the hazard reports. In the Amazon dataset, a total of 5,499 sentences (extracted from the 740 reviews with verified safety concerns) were manually tagged for each hazard type (see Chapter 1, again). For instance, the categorical (tag) type "Hazard Type" may consist of the code types "Fall," "Mechanical," "Suffocation," and so on.

According to the tagging protocol (adapted from Chapter 1), both tagger A and an authoritative tagger, tagger B, individually tagged a random set of focal sentences for each code value in each safety concern category. For the Amazon dataset, authoritative tagger B tagged all 5,499 sentences for classifying the code values of the hazard type. Then, tagger A tagged the random sets for classifying the code values, as necessary, for validation purposes. First, tagger A tagged a small portion to check whether the protocol was clear enough to

³⁰Health Canada. *Health Canada Decision-Making Framework for Identifying, Assessing, and Managing Health Risks*. Retrieved from: <https://www.canada.ca/en/health-canada/corporate/about-health-canada/reports-publications/health-products-food-branch/health-canada-decision-making-framework-identifying-assessing-managing-health-risks.html>

³¹Union, E. Guidelines for the management of the European Union Rapid Information System 'RAPEX'. Retrieved from: https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/repository/content/pages/rapex/docs/Guidelines%20annex_en.pdf

follow the tagging instructions. To verify the interrater agreement (i.e., reliability) between taggers A and B, we computed Cohen's kappa for each code value, which indicated a substantial agreement (Landis and Koch, 1977). After tagger A tagged the first set of 200 focal sentences from the Amazon dataset to classify each code value of the hazard type code category, the kappa score was 0.82. After tagging the last set of 200 focal sentences from the Amazon dataset, the score was 0.85 for the hazard type. Hence, there was an improvement in the kappa scores because the taggers resolved the discrepancies in their tags. After resolving the conflicts, we kept the tags with the agreement between taggers A and B for the smoke term generation.

4.4. Generation of smoke words

Using the Amazon.com dataset, the researchers performed significant word analysis for identifying what phrases (n-grams or smoke words) were present in the individual code types. We ran a machine-learning algorithm (automated approach), which uses the prevalence measure or correlation coefficient (CC) score and was developed by Fan, Gordon, and Pathak (2005).

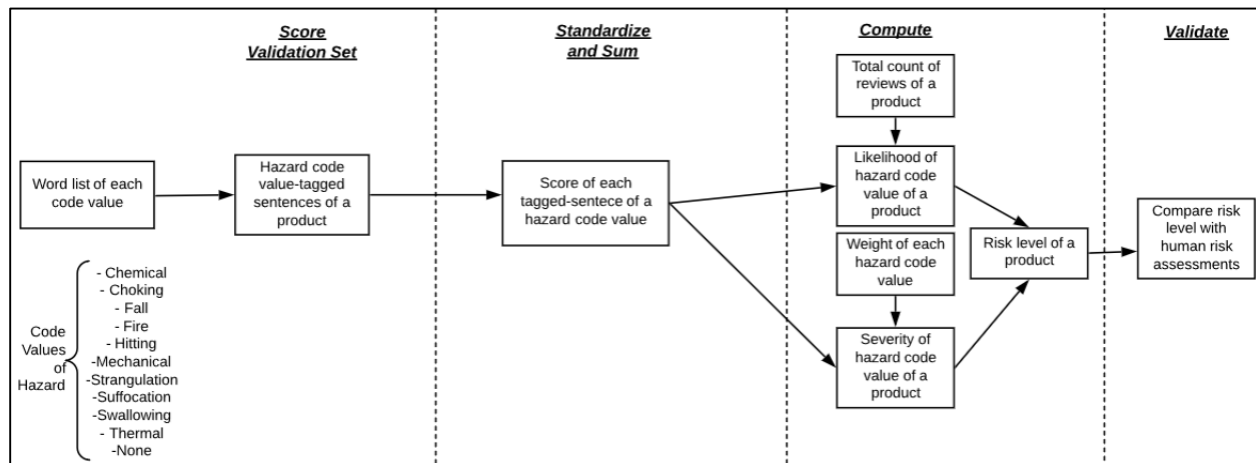
First, the reviews tagged with each code value were divided to include 80 percent in a training set and 20 percent in a validation set. From the training set, this automated approach rapidly identifies words that are more prevalent in a specific code value (e.g., chemical hazard) than in other code value (e.g., no chemical hazard). After employing the machine-learning algorithm (Fan, Gordon, and Pathak, 2005) to extract initial candidate smoke terms, we manually curated the terms to improve generalizability. According to the prior studies

(Abrahams et al., 2012; Winkler et al., 2016), we removed common English words (e.g., and, but, it) and removed words containing specific brand names. Additionally, we removed words with CC scores less than or equal to 0. At the final stage, words with high prevalence in a specific code value are stored in the “n-gram dictionary” along with the CC scores assigned to each prevalent word.

4.5. Risk scoring procedure

Figure 2 illustrates how the word generation of each hazard code value was applied to our risk-assessment scoring model for computing the likelihood and severity of a hazard code value. The results obtained from computing the specific hazard likelihood and severity were incorporated into determining the overall risk level of a specific product.

Figure 2. Risk scoring procedure



As delineated in the Formula (1) below, the word list of each hazard code value was used to score each focal sentence of a product from the validation set. For computing the standardized score of each focal sentence, we divided the score of each focal sentence by the maximum smoke score sentence for that particular product. This sentence smoke score standardization

procedure was only used if the maximum product smoke score exceeded 10,000 and if no sentence score exceeded that threshold, all standardized sentence smoke scores were set to zero. Then the scored sentences were manually reviewed for detecting the false positives (i.e. positive scored-sentence that did not mention a true hazard incident were discounted). We set the standardized sentence smoke scores to zero for the false positives that we found during the manual review. Though this step introduced some manual intervention into the procedure, the overall effort is modest, since comparatively few sentences from the overall dataset need to be manually reviewed. The payoff from additional manual review is substantial, as false positives that may bias the risk assessment scores can be weeded out. After discovering the scored sentence of the true positive, s , we then took the summation of the standardized scores of the true positives, S , of the specific hazard code value, c , of a product, p , to arrive at a standardized product smoke score $S_{c,p}$.

$$S_{c,p} = \sum_{i=1}^n s_i \dots\dots\dots \text{Formula (1)}$$

n = total number of scored-sentences that are true positives of a specific hazard code type, c , of a product, p .

For example, $n = 2$, then $S_{c,p} = s_1 + s_2$. That is, if there are two hazard-mentioning sentences (s_1 and s_2), each with a standardized smoke score, for a particular product the standardized smoke scores of each of the two sentences can be added with simple addition, to obtain a standardized total smoke score, for the product, where the standardized total smoke score is denoted $S_{c,p}$.

4.5.1. Estimating Hazard Likelihood for a product

To estimate the hazard likelihood, we mapped the count of Amazon reviews – which is a proxy for product sales (Liu et al. 2017) and hence may approximate product penetration – to the EU Risk Likelihood scale. To implement this mapping, and approximate a risk likelihood, we took the standardized smoke score ($S_{c,p}$ defined above) of reviews mentioning a hazard of a specific type (c) for the product (p) on Amazon.com, and divided by the total count of customer reviews, R, for that product (p) from Amazon.com:

$$\text{Likelihood of a hazard code type} = S_{c,p} / R_p \dots\dots\dots \text{Formula (2)}$$

4.5.2. Estimating Hazard Severity for a Product

To compute the severity level of the hazard code type of a product, we first considered the count of sentences mentioning a specific hazard code type across all product reviews for that product model. As shown in Table 6, we first count the number of sentences mentioning the specific hazard code type of a product, and then assign a weight, w_s ,

which escalates as the number of sentences mentioning a hazard increases, but plateaus when this count reaches 3 (since additional mentions likely have no marginal utility).

Count of sentence(s) mentioning a hazard code type	Assigned weight, w_s
0	0
1	0.33
2	0.67
3 or more	1

Table 6: Assigned Weight of Sentences Mentioning a Hazard Code Type

We combined the weight, w_c , of risk for each hazard code type with the assigned weight of the sentence(s) mentioning the hazard code type of a product to compute the hazard severity level of the product:

$$\text{Product Hazard Severity level, } SL_p = (w_s) * (w_c) \dots\dots\dots \text{Formula (3)}$$

4.5.3. Estimating Severity for a Hazard Type

We assumed that the risk of each hazard code type (w_c) would be weighted differently. For example, for the code value of a hazard category, a fire hazard could receive a higher weight than the chemical hazard did. In contrast, a fire hazard could have a lower weight than a strangulation hazard would. We surveyed product safety experts, who then manually decided the weight of risk for each hazard code value. We emailed 54 experts (i.e. survey participants) through Qualtrics to invite them to participate in the survey. The survey participants were identified from attendees at a major international event for product safety experts: the 2019 Annual Symposium of the International Consumer Product Health and Safety Organizations (ICPHSO). A small subset of attendees was solicited to participate, and snowball sampling was used (participants were asked to

recommend colleagues). 5 out of 54 respondents completed the survey. Table 7 shows the total time (in hours) taken by each respondent in completing the survey along with start and end dates.

Respondent ID	1	2	3	4	5
Start date	04/05/19	04/06/19	04/05/19	04/08/19	04/09/19
End date	04/05/19	04/06/19	04/07/19	04/09/19	04/09/19
Total time (hour)	1.2	2.4	48.3	23.0	3.3

Table 7. Time taken for completing the survey by human experts

The average survey time, 15.6 hours, was higher than the expected time, 1.5 hours to complete the survey. The survey software was unable to track rest-times taken by participants. Nevertheless, the overall survey duration seems to indicate all participants carefully completed the survey. Furthermore, the thoughtful risk rationales captured by each human expert (See Appendix B), substantiate that experts were diligent in their assessments. We note from the individual survey durations that the time taken by human experts to assess the risk of 10 products is substantially more than the 1.5 hours we originally estimated when we piloted the survey internally ourselves (5-10 minutes per product). This implies both (1) that the time-saving from automated product risk assessment could be higher than we had conservatively anticipated, and (2) that, in future surveys, fewer products should be included per survey.

Table 8 shows the experts' years of experience in product safety and product risk assessment, for the 5 respondents.

Field	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Mean	Min	Max
Product safety	15	9	27	26	26	20.6	9	27
Product risk	15	9	17	26	26	18.6	9	26

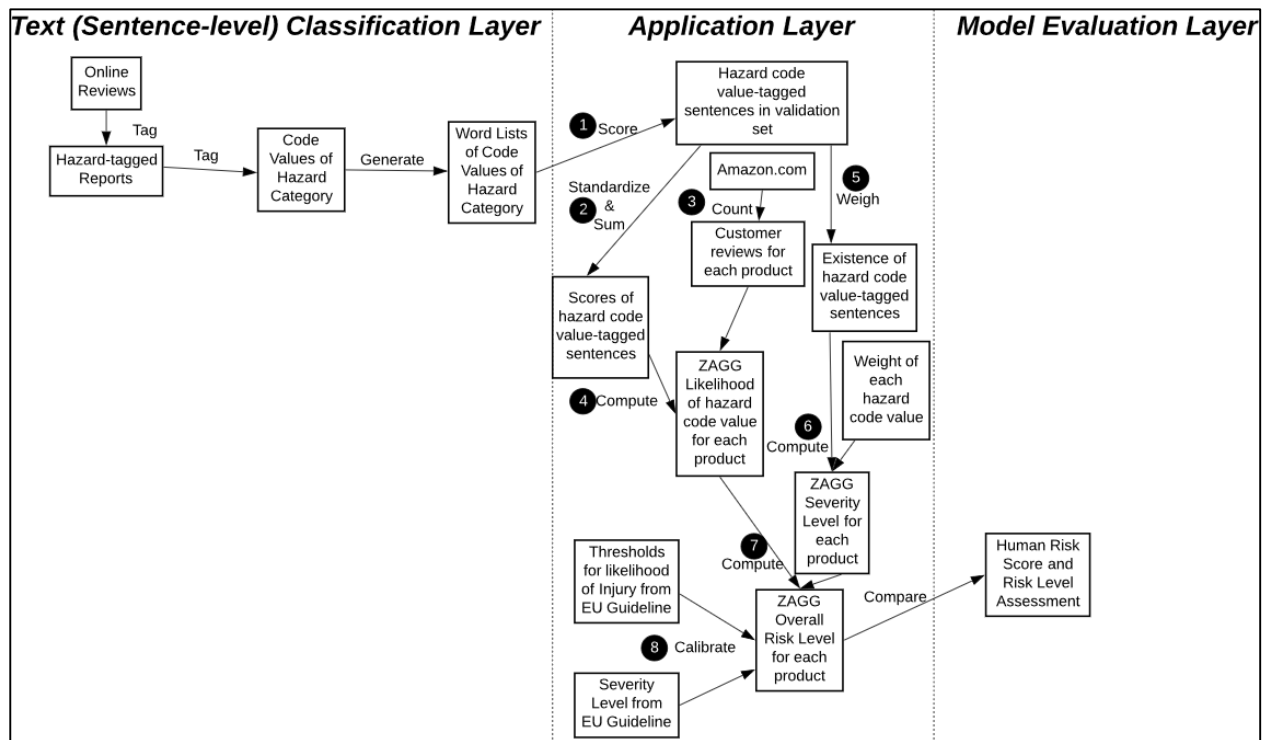
Table 8. Experts' years of experience in product safety and product risk assessment

Table 8 shows us that the average number of years of experience in product safety and product risk assessment for our respondents was close to 21 years and 19 years, respectively. We believe the hazard weights solicited from this small group of highly experienced experts holds greater validity than collecting a large set of human assessment results from non-experts. To determine severity weights for each hazard code value, we simply averaged the weights assigned by each of the 5 respondents, for that hazard code value.

4.6. Procedure for evaluating the risk assessment model

Once the online reviews were classified into hazard categories, the risk model was then defined and applied, and finally, the risk model was evaluated, as shown in Figure 3.

Figure 3. Overall study methodology: defining and evaluating the risk-assessment model



The risk levels and coding schemes for the hazard code types were adapted from the coding manuals and risk assessment frameworks of Health Canada and the EU guidelines. Our model first included the automated approach in generating the words that were most prevalent in each code type. *Step (1)*: The automated approach of scoring the hazard-specific sentences was incorporated into our risk assessment model to score the sentences for specifying the hazard code type (see Figure 2). *Step (2)*: The score of each sentence was standardized to obtain a value, scaled from 0 to 1. By using the Formula (1), we computed the summation of all the standardized scores with sentences that mentioned an incident specific to

a hazard code type and a sentence-smoke score having a threshold value more than 10,000.

Step (3): Then we included the raw count of online reviews of a particular product from

Amazon.com. *Step (4):* By using the Formula (2), we computed the likelihood of hazard code

type. *Step (5):* Referring to the assigned weights, w_s shown in Table 6, we used the existence

of hazard-specific sentences by counting the number of sentences mentioning the specific

hazard code type of a product. *Step (6):* By using the Formula (3), we combined the weight of

the risk for each hazard code type, w_c with the assigned weight of the sentence(s) mentioning

the hazard code type of a product, w_s to compute the hazard severity level of the product, SL_p .

Step (7): Therefore, the model measured two different risk scores—the likelihood of the specific hazard code type and severity level of a hazard—for determining the risk level

(ranging from low to serious) of a product. This study adapted the risk matrix chart from the

EU guideline to locate the risk level of the product (see Figure 4).

Figure 4. Risk matrix chart (adapted from the EU guideline)

Probability of damage during the foreseeable lifetime of the product		Severity of injury			
		1	2	3	4
High Low	> 50 %	H	S	S	S
	> 1/10	M	S	S	S
	> 1/100	M	S	S	S
	> 1/1000	L	H	S	S
	> 1/10000	L	M	H	S
	> 1/100000	L	L	M	H
	> 1/1000000	L	L	L	M
	< 1/1000000	L	L	L	L

S – Serious risk
H – High risk
M – Medium risk
L – Low risk

Step (8): To calibrate our risk scores with the risk matrix of the EU guideline, we scaled the four categorical hazard severity levels into numeric scores from 0 to 1 (as shown in Table 9) and scaled the likelihood of hazard code type from 0 to 1.

EU guideline's Severity level (EU)	Corresponding ZAGG numeric severity level (s)
1	$0 < s \leq .25$
2	$.25 < s \leq .50$
3	$.50 < s \leq .75$
4	$.75 < s \leq 1.0$

Table 9. Mapping EU guideline's Severity level (EU) to ZAGG Severity level (s)

Our risk matrix shows that the severity level was grouped into four blocks to calibrate the 4 levels of severity adapted from the risk matrix of the EU guideline (see Figure 4). Then the combination of the likelihood values and the severity levels (adapted from EU guideline) computed in our computed risk scores is used to determine the product's *overall* risk level (Low, Medium, High, or Serious), on the grid.

4.7. Risk model assessment with the computer versus human experts

We used the Pearson correlation between the computer algorithm–derived score and reconciled expert (human) score for each risk metric. We reported separate Pearson correlations for the likelihood of hazard and hazard severity. Getting R close to 1 will show high correlation/agreement between the computer score and human expert score.

We performed the Pearson correlation on comparing the hazard severity scores between computer versus human and human versus human judgments, as well as comparing the likelihood of hazard between computer versus human and human versus human decisions.

These tests can confirm whether humans are consistent and reliable on the risk assessment of a specific product, as well as whether the computer scoring method is reliable for assessing a product's risk level.

5. Results and Discussion

The following subsections report our results in each phase of model building and assessment:

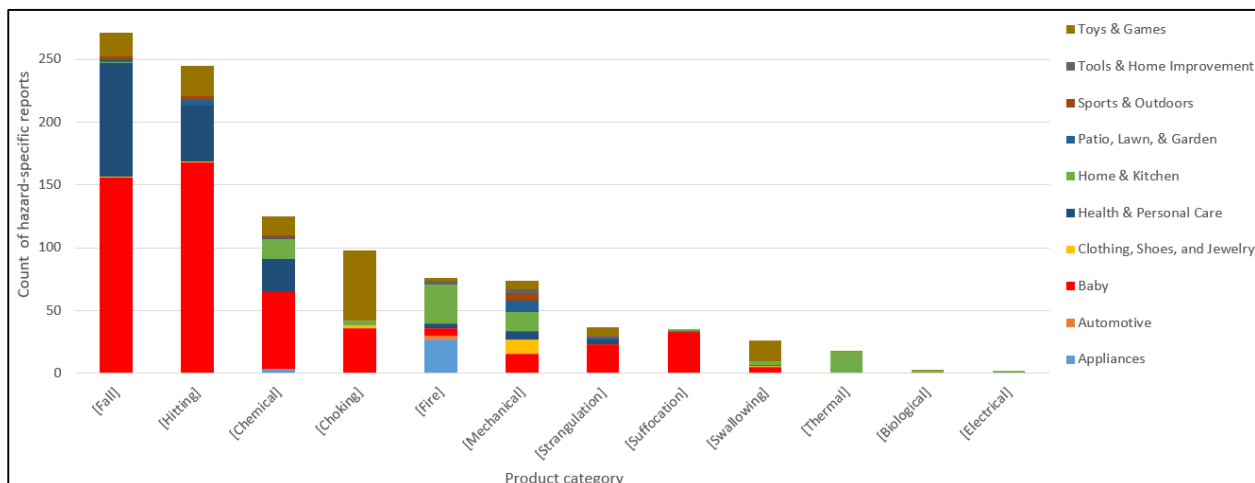
5.1. Scoring Products

Each product review in the Amazon.com was tagged for the existence of a safety concern.

The safety concern in our study is primarily focused on discovering whether a hazard incident exists or not exists across the multiple product reviews (i.e. 740 hazard-tagged reviews) that contain 5,499 focal sentences.

All the hazard code types mentioned in the chart above were obtained through manual tagging (see Figure 5). Figure 5 shows various code values of the hazard category mostly present among baby products.

Figure 5. Count of sentences with hazard existence, by hazard code type and product category



A high count of sentences mentioned the fall hazard code type for baby and health and personal care products. The hazard reports for baby products frequently mentioned that babies and toddlers tend to fall, and this mostly occurs from toddler beds, booster seats, and car seats. Compared with other product categories, there is also a high count of reports of health and personal care products mentioning the fall code value, as the reviewers mentioned the tendency to fall when using products like wheelchairs, walking sticks, and toilet safety rails. This tells us that there are reports mentioning specific code values of hazards in the sentences of online reviews from customers. Hence, with the help of our risk-assessment scoring model, generating the distinctive words that are most prevalent in each hazard code type and then scoring the hazard-specific reports would benefit the stakeholders by locating the overall risk level of a product from a large pool of online consumer-generated reviews.

Table 11 gives an example of the risk-score computation of a low-risk consumer product, G, and a high-risk consumer product, H, identified by our computerized risk assessment model.

Product	Sentence ID number	Step 1: Unigram score	Step 2: Standard score	True/False Positive (score > 10000)	True standard score	Step 2: Cumulative standardized score	Step 3: Count of customer reviews	Step 4: Likelihood of hazard code type	Step 5: w_s : Existence of hazard-specific sentences	$\bar{w}_c$: Mean weight of hazard code type	Step 6: Severity Level of Hazard	Step 7: Overall Risk Level
Low-risk product: G	1	21368.9	1	True	1	1	1228	1/1228 = 0.0008	0.33	0.6	0.33*0.6 = 0.20	Using Risk Matrix adapted from EU Guideline (see Table 14), Risk Level = Low , when Likelihood = 0.0008 and Severity = 0.20.
	2	9587.4	0.45	False	0							
	3	6428.7	0.30	False	0							
	4	5492.2	0.26	False	0							
High-risk product: H	1	22233.7	1	True	1	1.50	466	1.50/466 = 0.003	0.67	0.6	0.67*0.6 = 0.40	Using Risk Matrix adapted from EU Guideline (see Table 14), Risk Level = High , when Likelihood = 0.003 and Severity = 0.40.
	2	20006.4	0.89	False	0							
	3	12186.1	0.55	False	0							
	4	11508.1	0.52	False	0							
	5	10480.3	0.50	True	0.50							

Table 11: Computation of the Risk Scores of a **Low**-risk and **High**-risk Product

Step (1): We displayed the scored sentences of low-risk product, G by scoring the sentences with the fall hazard's one-word (unigram) list. After manually reviewing the sentences, we noted the true positives (Column 5 of Table 11). Table 12 shows the total score of the sentences that we noted as true positives of the fall hazard reports of low-risk and high-risk products, G and H.

Product	Sentence #	True positive of fall hazard report	Total score
G	1	If my granddaughter leans ^{6349.7} forward, the whole ^{5552.6} chair ^{454.6} tips ^{9012.1} .	21368.9
H	1	I grabbed ^{4191.5} this so-called safety bar ^{4649.7} which was correctly attached to my shower wall, next thing I knew, the suction cups detached from the wall and I was waving the stupid thing in the air ^{3393.1} while catching ^{1047.5} myself from falling ^{8951.9} .	22233.7
H	5	Had her helper not been with her she would have fallen ^{10480.3} and certainly injured herself seriously.	10480.3

Table 12. Scoring procedure of true fall hazard reports

Each sentence indicating a true positive of a fall hazard is shown in Table 10. The smoke scores for each word that was matched in the fall hazard smoke unigram dictionary is shown as a superscript above of each bolded unigram in Table 12. As a result, the summation of these unigram smoke scores contributed to the overall sentence score (e.g. $6349.7 + 5552.6 + 454.6 = 21368.9$). Then, after following the same procedure, we displayed the true positives (see Table 12) specific to the fall hazard incident of the high-risk consumer product H after scoring these sentences with the fall hazard's 1-word (unigram) list. *Step (2):* After obtaining the standardized scores of the true positives for each product, we calculated the cumulative scores. *Step (3):* Then we collected the raw counts of online reviews: 766 and 466 of products G and H, respectively from Amazon.com. *Step (4):* By using the Formula (2), we computed the likelihood of fall hazard of each product. For product G, the likelihood of fall hazard was 0.0008 and for product H, the likelihood was 0.003. *Step (5):* To calculate w_s for product G, we counted one sentence that mentioned the incident of a fall hazard of product G. Hence, w_s

= 0.33. After following the same procedure, we discovered 2 sentences that mentioned the incident of a fall hazard of product H. Then, $w_s = 0.67$. *Step (6)*: By using the Formula (3), we combined the mean weight of the risk for fall hazard, $\bar{w}_c$ (assigned by the human experts) with the assigned weight of the sentence(s) mentioning the fall hazard of each product, w_s to compute the hazard severity level of the product, SL_p . Therefore, for product G, $SL_G = 0.17$ and for product H, $SL_H = 0.34$.

5.2. Severity Weights for Hazard Codes

As we assigned weight to each hazard type before the survey, we compared our results (i.e. pre-survey weight, w_c) with the weights assigned by 5 experts on the survey (see Table 13).

Hazard code type	w_c	w_1	w_2	w_3	w_4	w_5	Min	Max	Q_1	Q_3	Mean, $\bar{w}_c$	Standard Deviation
Chemical	0.2	0.2	0.3	0.5	0.5	1	0.2	1	0.3	0.8	0.5	0.3
Choking	0.5	0.8	1	1	1	1	0.8	1	1.0	1.0	1.0	0.1
Mechanical	0.5	0.1	0.2	0.5	0.3	1	0.1	1	0.2	0.8	0.4	0.4
Fall	0.5	0.5	0.8	0.3	0.3	1	0.3	1	0.3	0.9	0.6	0.3
Fire	0.5	0.5	1	0.9	1	1	0.5	1	0.9	1.0	0.9	0.2
Hitting	0.2	0.1	0.5	0.3	0.3	0.5	0.1	0.5	0.3	0.5	0.3	0.2
Strangulation	0.5	1	0.5	1	1	1	0.5	1	1.0	1.0	0.9	0.2
Suffocation	0.5	1	1	1	1	1	1	1	1.0	1.0	1.0	0.0
Swallowing	0.6	0.5	0.4	0.3	1	0.7	0.3	1	0.4	0.9	0.6	0.3
Thermal	0.4	0.2	0.2	0.5	0.5	0.7	0.2	0.7	0.2	0.6	0.4	0.2

Table 13. Experts' assigned weights to each hazard code type

Table 13 shows the weights assigned by each expert represented as w_1 , w_2 , w_3 , w_4 , and w_5 . In order to detect outlier(s) of each weight, we calculated the first quartile, Q_1 and the third quartile, Q_3 of the experts' assigned weights. As we did not find outliers for any one of the weights assigned by the experts, we, therefore, incorporated the simple mean weight (average weight of all five experts, for that hazard type) of each hazard code type into our risk

assessment model to compute the severity level of hazard of each product by using Formula (3).

5.3. Positioning Products on the EU Risk Matrix

Step (7): After performing the calculations of the risk scores for the low-risk product G and high-risk product H, we mapped the likelihood and severity level of a fall hazard for locating the *overall* risk level of each product. For the low-risk product G, the ZAGG likelihood of a fall hazard was 0.0008 and the ZAGG severity level of the hazard was 0.20. For the high-risk product H, the ZAGG likelihood of a fall hazard was 0.003 and the ZAGG severity level of the hazard was 0.40.

Step (8): As a result, the risk matrix indicated a Low risk level for product G because the likelihood of a fall hazard was greater than 0.0001 and the ZAGG severity level was in the range of 0–0.25, corresponding to the severity level 1 according to EU guideline’s risk assessment matrix. In contrast, the risk matrix (see Table 14) labeled a High risk level for product H because the likelihood of a fall hazard for this product was greater than 0.001 and the ZAGG severity level was in the range of 0.25–0.50, which corresponds to the severity level 2 according to the EU guideline’s risk assessment matrix. Hence, our study adapts the risk-level chart in Figure 4, from the international standards of the EU guidelines, to build the risk-level matrix (see Table 14)

Likelihood of hazard code type	Severity level (EU) from EU guideline and a range of computerized severity level (s) of the hazard code type			
	EU 1	EU 2	EU 3	EU 4
	$0 < s \leq .25$	$.25 < s \leq .50$	$.50 < s \leq .75$	$.75 < s \leq 1.0$
> .50				
> .10				
> .01				
> .001		Product H		
> .0001	Product G			
> .00001				
> .000001				
< .0000001				

Risk level	Response
Serious risk	<i>High priority investigation</i>
High risk	<i>Immediate investigation</i>
Medium risk	<i>Investigate</i>
Low risk	<i>Take no action</i>

Table 14: Mapping Computerized Risk Level on the Risk Assessment Matrix (Adapted from the EU Guideline)

By looking at the risk location of product H, the safety regulator may need to step in to take action for the “High Risk” level through a high-priority investigation of product H. In contrast, the risk location of product G guides the safety regulator to take no current action on product G as its overall risk level is “Low Risk”.

We identified the risk levels of various products across multiple hazard code types (chemical, choking, fall, fire, etc.) after computing the risk scores (i.e., likelihood and severity level) and then mapping these computed risk scores on the risk matrix chart adapted from the EU guideline. Based on locating the specific risk level of a product, a safety regulator can respond to take proper action (as listed under Table 14).

We selected 14 random low (L) overall risk products using the ZAGG (computer) model and 16 random products with Medium (M) to Serious (S) overall risk using the ZAGG

(computer) model. These 30 products were assigned at random to the 5 survey respondents, to elicit their opinions as to each products hazard likelihood, hazard severity, and overall risk. Each expert was presented with 1 product (for training) followed by 10 random products for assessment. Random subsamples of 10 random products for each expert were chosen to reduce assessment fatigue, as complete assessments for all 30 products would have been excessively burdensome for individual experts. Each expert has first asked to assess each product's hazard likelihood, severity, and overall risk by first consulting only the raw Amazon reviews for that product, directly from Amazon.com. Then, the expert was presented with specific hazard narratives for the particular product, that were highlighted by our smoke-scoring procedure, and asked to re-assess the product's hazard likelihood, severity, and overall, given the specific narratives highlighted by us.

We selected 14 Low (L) overall risk products using the ZAGG (computer) model - humans agreed in 11 out of 24 products (i.e. 46% of post-judgments) that the product was Low (L) overall risk (see Table 15). (Note there are more product assessments than products, as some products were presented to multiple experts, due to the randomization procedure for selecting products for presentation to experts.) Table 15 assesses the agreement between the pre-calibrated ZAGG-computed risk level and the risk level assessed by the human experts. These *pre*-calibrated ZAGG risk levels and their pre-calibrated ZAGG risk scores were determined by using the *default* weights of the hazard codes, w_c .

Risk level	Count of products	Experts' agreement	Count of evaluations	Percentage in agreement
Low (L)	14	11	24	45.8%
Medium (M)	6	1	15	6.7%
High (H)	4	0	6	0.0%
Serious (S)	7	8	17	47.1%

Table 15. Assessing the agreement between *Pre*-calibrated ZAGG-computed risk level with *default* weights of hazard codes and human expert-assessed risk level

For instance, when we found 7 products with Serious (S) overall risk by using the pre-calibrated ZAGG risk model, we observed the experts agreed in 8 out of 17 evaluations that the product had a Serious (S) overall risk. (Again, there were more evaluations than products, as each product may have been presented to multiple experts, due to the randomization procedure.) Therefore, the experts agreed 47.1% of the time with the pre-calibrated ZAGG-computed risk model that a product had a Serious (S) overall risk.

Table 16 shows the revised, *post*-calibration ZAGG scores, that were computed using the experts' average weight of each hazard code, $\bar{w}_c$ as determined *after* the expert survey (rather than the default, pre-calibration weights). We then compared the agreement between the post-calibrated ZAGG-computed risk level and the risk level assessed by the human experts.

Risk level	Count of products	Experts' agreement	Count of evaluations	Percentage in agreement
Low (L)	15	12	26	46.2%
Medium (M)	3	0	6	0.0%
High (H)	4	1	7	14.3%
Serious (S)	9	9	12	75.0%

Table 16. Assessing the agreement between *Post*-calibrated ZAGG-computed risk level using *experts' average weight* of each hazard code and human expert-assessed risk level

For example, when we discovered 9 products with Serious (S) overall risk by using the post-calibrated ZAGG risk model, we found that the experts agreed in 9 out of 12 evaluations that the product had a Serious (S) overall risk. (There were more evaluations than products, as

each product may have been distributed to multiple experts, due to the randomization procedure.) Therefore, the experts agreed 75.0% of the time with the post-calibrated ZAGG-computed risk model that a product had a Serious (S) overall risk. After using the expert's average weight of each hazard code, we observed an improvement (47.1% to 75%) in experts' agreement with our post-calibrated ZAGG risk level in assessing the Serious (S) overall risk. On the other hand, there was a drop in agreement (6.7% to 0.0%) between the experts' Medium (M) overall risk and the post-calibrated ZAGG overall risk level.

Table 17 shows the pre-calibrated ZAGG-computed risk levels of each product using the default hazard code weights, w_c . Also, Table 17 displays the post-calibrated ZAGG-computed risk levels of each product using experts' average weight of each hazard code, $\bar{w}_c$. Additionally, Table 17 shows the 5 human experts' assessments on the likelihood and severity scores of the hazards, for each of the 30 products in our survey. For comparing the ZAGG (computer-assessed) scores to the experts' judgments, Table 17 displays the average scores of each risk metric (i.e. likelihood and severity) when the experts had only read the product reviews directly from Amazon and then also shows the average scores after the experts assessed the risk by reading the specific hazard narratives for the particular product, that were highlighted by our computerized smoke-scoring procedure.

Hazard code value	Product number §	Total count of reviews from Amazon	Likelihood			Severity			Risk level				Count of response(s)	
			ZAGG* (Pre-calibrated)	Mean Human Expert Before**	Mean Human Expert After***	ZAGG* (Pre-calibrated)	Mean Human Expert Before**	Mean Human Expert After***	ZAGG* (Pre-calibrated)	ZAGG**** (Post-calibrated)	Human Expert Before**	Human Expert After***	Before	After
Chemical	28	3	0 (0)	0.001	0.001	0.13	0.5	0.5	L	L	S/M	S/M	2	2
	21	51	0.06 (4)	0.002	0.002	0.003	0	0	M	S††	L/L/M/L/M	L/L/M/L/M	5	5
	22	2	0.5 (1)	0.1	0.1	0.86	0	0	S	H††	M/L	M/L	2	2
	24	1	1 (1)	0.0000001	0.0000001	.07	0	0	H	H	L/H	L/S†	2	2
Choking	1	766	0.002 (2)	0.01	0.01	0.5	0.75	0.75	S	S	S/S	S/S	2	2
	3	2	0.9 (2)	0.004	0.01†	0.34	0.25	0.75†	S	S	S/L	S/S†	2	2
	5	273	0 (0)	0.000001	0.000001	0	0	0	L	L	L/M	L/M	2	2
	7	2	0.58 (2)	0.5	0.5	0.34	0.75	0.75	S	S	S	S	1	1
	9	23	0 (0)	0.00001	0.00001	0	0.25	0.25	L	L	L	L	1	1
	14	360	0.003 (1)	0.0001	0.0001	0.17	0.5	0.5	L	H††	H	H	1	1
Fall	18	1811	0.0006 (1)	0.0000001	0.0000001	0.17	0	0.25†	L	L	L/L	L/M†	2	2
	2	750	0.001 (1)	0.01	0.01	0.17	0.75	0.75	L	L	S	S	1	1
	6	1228	0.001 (1)	0.001	0.001	0.17	0.75	0.75	L	L	S	S	1	1
	25	1037	0.001 (2)	0.001	0.001	0.34	0.25	0.25	H	S††	M	M	1	1
Fire	26	466	0.003 (2)	0.4	0.1†	0.34	0	0	S	S	S/H	S/M†	2	2
	4	183	0 (0)	0.0000001	0.0000001	0	0	0	L	L	L	L	1	1
	10	64	0.016 (1)	0.0001	0.0001	0.17	0.25	0.25	M	L††	M	M	1	1
	11	3	0.33 (1)	0.0001	0.0001	0.17	0.75	0.5†	L	M††	H/S/S/L	H/S/L†/L	4	4
Hitting	19	87	0 (0)	0.000001	0.000001	0	0.25	0.5†	L	L	L/S/L	L/S/M†	3	3
	23	153	0 (0)	0.000001	0.000001	0	0.25	0.25	L	L	L/H	L/S†	2	2
	27	58	0.032 (2)	0.3	0.3	0.034	0.5	0.5	M	M	S/S	S/S	2	2
Mechanical	8	102	0.01 (1)	0.000001	0.000001	0.17	0	0	M	M	L	L	1	1
	18	1811	0.001 (2)	0.0000001	0.0001†	0.5	0	0.25†	H	S††	L/L/L	L/M†/M†	3	3
	25	1037	0 (0)	0.000001	0.00001†	0	0	0	L	L	L	L	1	1
Strangulation	15	2	0 (0)	0.000001	0.000001	0	0	0	L	L	L	L	1	1
	27	58	0.017 (1)	0.3	0.3	0.07	0.5	0.5	M	S††	S/S	S/S	2	2
Suffocation	16	52	0.0014 (2)	0.0000001	0.0000001	0.13	0.75	0.75	M	S††	L/L	L/L	2	2
	17	727	0.034 (0)	0.5	0.5	0	0	0	L	L	S/S	S/S	3	3
Swallowing	12	3	0.3 (1)	0.034	0.004†	0.2	0.25	0.5†	H	H	S/L	S/L	2	2
	29	393	0.003 (1)	0.000001	0.0000001	0.2	0.5	0.25†	L	L	L/L/L	L/L/L	3	3
Thermal	20	1527	0.001 (0)	0.00003	0.00003	0	0.25	0.25	L	L	L/H/L	L/H/L	3	3

Table 17: Risk Level of Each Product for Its Specific Hazard Code Value

§ contains 28 unique products that are anonymized by randomly assigning distinct product number to each product, i.e. each product contains at least one hazard code; *Pre-calibrated ZAGG-Computed risk score/level using default hazard code weights; ** Human Expert Before-average risk score is given by reading reviews from Amazon; ***Human Expert After-average risk score is given after reading hazard snippet; **** Post-calibrated ZAGG-Computed risk level using experts' average weight of each hazard code; () - a number of the hazard-specific snippet(s) of a product; † indicates the human experts, on average, changed their risk assessment score after being presented with hazard snippets; †† indicates the pre-calibration ZAGG-computed risk level changed when including the experts' average weight of each hazard code in the post-calibrated ZAGG risk-scoring model.

Out of 54 participants, 5 human experts fully responded to the online survey in assessing the hazard risk of 10 randomly-assigned products. As a result, the determination of risk level was not uniform across all the products (see Table 17). For instance, only 1 expert assessed the product number 7 in a random chance for determining the risk level: serious, “S” of chemical hazard when reading the reviews on Amazon and also recorded the same risk level after reading the specific hazard snippets of the product. On the other hand, all 5 human experts randomly received the survey question in assessing the risk levels of product number 21. Hence, we received a series of risk levels: Low (L)/Low (L)/Medium (M)/Low (L)/Low (L) from the judgments of 5 experts, for product number 21.

Table 17 shows a controversial result regarding determining the risk level by the computer versus human experts. For example, one expert determined a serious hazard risk level (S) of product number 6. Whereas, our risk-assessment-scoring model determined a low-risk level (L) of the same product after computing a low likelihood score of fall hazard, 0.001 along with a low severity score, 0.17. Obtaining a low hazard risk level of this particular product is due to discovering only one hazard snippet by our ZAGG-risk model. Contrary to the ZAGG risk assessment, the human expert’s rationale for the product’s risk assessment is given under the following statement: *“There is a hazard from it being too easy for a toddler to climb over and fall; there was one report of a child landing on his head after scaling the gate. There was also at least one report of a part suddenly breaking loose and shooting into the air, striking a child”*. Both the computer and the human expert scored the same likelihood of fall hazard, 0.001. However, the statement given by the expert confirms to us that the expert gave a heavier weight (0.75) to the hazard severity of this specific product than the computer discovered. Our ZAGG-computed risk methodology was successful in rapidly detecting these

high-scored sentences prevalent in hazard-specific narratives as the human expert gained knowledge of the fall hazard incidents after reading the particular hazard snippets of product number 6. There are multiple rationales posted by the experts after assessing the risk for each product with respect to the hazard-specific code value (see Appendix B).

5.4. Risk Model Evaluation Results

To evaluate the validity of our risk assessment model, we used the Pearson correlation between the computer algorithm–derived score and the average (human) score of the experts for each risk metric. We conducted Pearson correlation to compare the likelihood of hazard between the pre- and post-calibration computer (ZAGG) scores versus the human judgements of risk scores when reading only raw Amazon reviews (“Human Expert before”) vs. after reading computer-discovered hazard snippets (“Human Expert after”) – see Table 18(a)³². We also compared the human risk likelihood and severity scores before vs. after reading the hazard snippets – See Table 18(b).

Risk metric	ZAGG vs. Human Expert before	ZAGG vs. Human Expert after
Likelihood	0.17	0.21
Severity (pre-calibrated)	-0.09	-0.02
Severity (post-calibrated)	0.06	0.19

Table 18 (a). Pearson correlations of ZAGG scoring method with human judgment on risk scores

Risk metric	Human Expert before vs. Human Expert after
Likelihood	0.94
Severity	0.88

³² Only risk severity is shown pre- vs post-calibration, as calibration does not alter the risk likelihood score.

Table 18 (b). Pearson correlations of human judgments of risk scores when reading raw Amazon reviews (“Human Expert before”) vs after reading computer-discovered hazard snippets (“Human Expert after”)

The Pearson correlations between pre-calibrated ZAGG-computed likelihood and the expert-assessed likelihood scores before reading and after reading the hazard-specific narratives were 0.17 and 0.21, respectively. Getting a low correlation (0.1 to 0.3) indicates a weak correlation (agreement) (Soper et al., 1917) between the computer score and human expert score. The low correlations of the pre-calibrated ZAGG scores and human scores of severity (-0.09 and -0.02) confirm the computer scoring method is limited for assessing a product’s risk level.

Conversely, the Pearson correlations between post-calibrated ZAGG-computed severity and the expert-assessed severity level before reading and after reading the hazard-specific narratives were 0.06 and 0.19, respectively. After assigning experts’ average weight of the hazard code to the post-calibrated ZAGG risk model, we observed an overall improvement regarding the agreement in assessing the hazard severity of a product with the human expert’s judgement. For instance, an increase in the correlation (from -0.02 to 0.19) indicates a boost in agreement of determining the hazard severity level between the post-calibrated ZAGG model and the human expert’s judgement after reading the hazard-specific snippets of a product. Even though there is an increase in correlation, the low correlations of the post-calibrated ZAGG scores and human scores of severity (0.06 and 0.19) indicate that the computer scoring method is still restricted in assessing a product’s risk level.

The statistical test implies that human (expert) judgment should be relied on for assessing hazard severity of a product. However, our computerized method for detecting hazard-specific sentences from a large pool of data caused human experts to frequently alter

their overall severity scores. As indicated by the † superscripts in Table 17, our results indicate that the experts occasionally, though rarely, change their decisions on hazard likelihood, hazard severity, and overall risk level after reading the hazard-specific sentences presented by the computerized method.

6. Limitations and Future Work

The following subsections describe limitations of our research, and suggest avenues for future work.

6.1. Sample Size and Selection Biases

Our sample was limited to 30 random products. In future work, we hope to expand the set of assessed products to a larger set, possibly of a few hundred products.

Regarding sample selection, our sample consisted of products with various levels of risk, according to our computer model. However, though some products had sentences with low smoke scores (below the 10,000 threshold), we neglected to select a control group of products with zero smoke scores for all sentences in their product reviews. This omission could have hindered our ability to observe stronger correlations between the computerized and manual risk assessment scores of products. Our survey also omitted very high risk products (e.g. products that were actually recalled), as Amazon.com removes listings for those products when the recall is announced. In future, we would like to maintain a real-time archive of Amazon product reviews, so that historic reviews can be presented to risk assessment experts even if the product listing has been removed from Amazon due to a recall.

6.2. Survey Participation

Our survey was hindered by low (<10%) participation, possibly due to the voluntary nature of the survey. Also, though 10 participants started the survey, only 5 completed the survey and signed the re-consent at the end of the survey, possibly due to fatigue in being administered a long (90 minute) survey, with no compensation. In future, we may provide a shorter survey (e.g. 5 random products, instead of 10), and/or offer remuneration to survey participants for their effort.

To preserve participant goodwill, we did not send reminders to the survey population after almost all the products in the pool had been assessed, resulting in reminders not being sent to approximately half of the population (about 30 participants). Once additional products are added to our pool, and the survey is shortened, we will encourage these experts again to participate, via an additional reminder. Alternatively, we will send all experts a new invitation offering a shorter survey, and compensation.

6.3. Accuracy in Risk Model Calibration

To improve the risk model calibration regarding the weights of a hazard code type (e.g., chemical hazard) with the experts' judgments, in future work, we will compute the principal eigenvector of the priorities for choosing the weights given by multiple experts. We will use analytic hierarchy process (AHP) analysis (Saaty, 2003) for developing the pairwise

comparison matrix with a view to reconciling the experts' various weights of a hazard code type and hence obtain the principal eigenvector.

The initial version of our ZAGG computerized risk assessment model only includes code values for hazard type, and weights for hazard types. Future versions should incorporate code values and code weights for other safety concern subcategories defined in Chapter 1.

Specifically, we need to include codes and code weights for event types (e.g., actual injury and treatment should be weighted more heavily than potential injury or pathway to injury or remediation solution) and injury types, and possibly also body parts affected (e.g. head could have higher weight than foot).

6.4. Objective Statistical Assessment of the Risk Model

For simplicity, we applied a simple agreement percentage to determine agreement between the categorical computer- vs. human- risk levels, and we applied a simple Pearson correlation to determine the association between the numeric computer- vs. human expert- risk scores. However, given observations are non-independent (multiple humans sometimes assess the same product), the assumption of independence is violated, and these statistical measures may understate the extent of agreement or association. In future work, we will apply more complex, but more appropriate, statistical tests, such as extended Kappa Statistics for categorical agreement – see (Pan et al., 2013) – or repeated measures correlation for numeric association (Bakdash, 2017) (e.g. using the `rcorr` package in R).

6.5. Mitigating Selection Biases

There are large discrepancies between some product categories. Due to the broad coding of some categories, there is bias in selecting the hazard types. For example, toy products may consist of hazard-tagged sentences of online reviews with more choking hazards than fall hazards, whereas products for elderly people may consist of reviews with more fall hazards than choking hazards. Due to the selection biases inherent in online data sources, it is vital for practitioners and stakeholders to be aware that the discovery of specific hazard code types is unlikely to be fully representative of the population experiencing the safety hazards of a certain product. Instead, the specific hazard code types should be viewed as supplementary to conventional data-collecting exercises, such as the CPSC's NEISS and OECD's GlobalRecalls datasets. The rapid discovery and investigation of the various classifications in this current study merely represent a target for emphasizing the risk assessment of hazards.

Further, our data sources were limited to a single major online retailer (Amazon). The effort to analyze real-time, unstructured data by manually monitoring the millions of textual reviews on third-party sites was prohibitive in this study, but should be attempted in future work if more resources become available.

6.6. Estimating Product Penetration More Accurately

In terms of obtaining the count of customer reviews from Amazon.com for a specific product model, we selected a distinct point of time to collect the number of reviews. For example, Amazon.com for a specific product model showed 65 customer reviews on January 20 and then 68 reviews on January 24, 2019. We chose to study one instance instead of considering

the variability of each count of these customer reviews for the specific product. According to a similar study regarding sales rank investigated by Brynjolfsson et al. (2003), in future work, we should check the customer reviews each day in a specified range of dates, for example, between January 30 and April 30, and then note any big changes in the counts. If there are any abnormal changes, such as increasing from 65 to 500, then we should report and justify this big jump in the number of customer reviews.

Our risk-assessment-scoring methodology is biased against products that are highly popular and only recently listed on Amazon, because the count of reviews that we used in our risk model will understate actual product penetration. Therefore, future work needs to overcome this recent listing bias by looking at review velocity (e.g., reviews per month, last month; reviews per year, last year) rather than considering only the total count of reviews, or alternatively, look at 3rd party product popularity data resources, such as the Nielsen product penetration data. Additionally, future work needs to investigate how to adapt the risk-assessment-scoring methodology to give a more appropriate risk score for outliers (products) with only 1 or 2 or 3 reviews present in Amazon.com.

6.7. Non-representativeness of the dataset

Another limitation is that only a small percentage of consumers post safety hazard reviews on third-party sites, which leaves a large majority of consumers out of the study. For example, there were only 740 unique hazard reviews found in the Amazon.com dataset in the study compared with the millions of people using the products. The lack of representativeness of the dataset means that our findings should not substitute for epidemiological studies of

product risk (such as injury prevalence estimation using the NEISS dataset). Our processes are intended to be *indicative* of safety concerns, not representative of safety concerns. Hence, our model should be regarded as complementary to conventional resources, and could potentially act as a “red flag” or “early warning system” for product safety issues.

We observed low correlations between the ZAGG-computed risk scores and the human expert-assessed risk scores in our study. We plan to include a random set of products in our future study that contains no hazard-specific reports to test and compare the risk assessment of risk-free products to that of high-risk products. Additionally, testing the risk scores of risk-free products will further test the efficiency of our ZAGG risk assessment model. Hence, the future study will avoid the bias in our model as we only determined the risk scores of the high- versus low-risk products in this study.

In order to enlarge our sample size of hazard reports, we plan to merge multiple data sources in addition to Amazon.com for our future study. For example, Walmart, Target, and eBay.com also contain a large pool of online reviews for consumer products. Executing the semi-automatic classification procedures on a sample size gathered from multiple sources will increase the chance of collecting a high number of unique hazard reports.

6.8. Further CPSC Strategic Goals

As this study focused on achieving a few goals listed by the CPSC strategic plan (see Table 2), we plan to establish new methodologies on accomplishing one or more remaining goals in future studies, for example, “Goal #1.4: Implement hazard surveillance of consumer products to discover recalled products”. Future work could develop the automatic text classification

approaches to detect recalled products that are being illegally sold in electronic marketplaces. As a result, safety regulators could rapidly identify sellers who are illegally selling recalled products online to consumers.

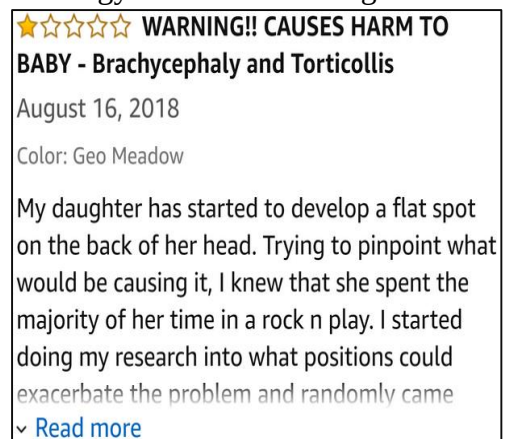
6.9. Classification of Injury Narratives

On April 12, 2019, the CPSC (2019a) recalled a Fisher-Price baby sleeper and also warned consumers to stop using the product if infants at 3 months or older begin rolling over in the product. The product was recalled due to reports of multiple infant deaths. This announcement motivates future study to update the text classification methodology to detect and categorize injury incidents from online reviews.

There are recent online reviews on Amazon posted on August 16, 2018, regarding injury incidents of the product. These consumer reviews mention narratives related to “death (unspecified cause)” and “harm (other cause)” code values that should be

included in classifying injury types in a future study. We could, for instance, add the following smoke terms:

- for “death (unspecified cause)” code value: “deaths”, “death of”, “SIDS”, “deaths in these”, “threat of SIDS”, “loss of her baby”, etc.
- for “harm (other cause)” code value: “causes harm”, “harm to”, “notified by my dr”, “the doctor”, “physical therapy”, “head and neck”, “neck issues”, “flat spot”, “back of her head”, “positions”, as well as formal diagnosis names like “torticollis” and “brachycephaly”.



7. Conclusion

In this study, we developed a model for determining the risk level of each product and exploring the text classification of hazard code types. We tested our risk assessment model after conducting the semi-automatic categorization of sentences using online reviews of products across multiple product categories. In addition, we included risk factors by adapting them from practitioners' risk-assessment frameworks and integrating online reviews with the text analytics to determine the computerized risk scores (i.e., likelihood of a hazard code type and severity level) and overall risk level of a product. Hence, these risk scores helped us map the risk level of a specific product on the risk-assessment matrix chart of the EU guideline. The semi-automated framework not only considered the content of the online (real-time) reviews but also utilized the gathering of raw counts of customer reviews from Amazon's online data source. Ultimately, the risk-level matrix may help support decision making for stakeholders like compliance officers, risk-management officers, and other risk-prevention experts; the framework may help them decide whether they should mitigate the risks or hazards related to specific product models.

8. Implications for Research and Practice

Our findings have implications for both research and practice. First, this study opens the opportunity for safety hazard researchers to build a knowledge hub of hazard narratives by shifting from privately used hospital datasets to using real-time (e.g., online) datasets. Second, we introduced the application of categorizing the real-time hazard narratives by assessing the hazard risk levels of various products. Third, we executed the risk-assessment methodology using the risk metrics adapted from the EU guidelines. In terms of the theoretical contribution in this study, safety hazard researchers can delve into extending "uncertainty theory" (Wu et al.,

2017), as we incorporated risk metrics like likelihood estimation, severity estimation, and hazard identification. These risk metrics were computed for building our risk-assessment scoring model for determining the hazard risk level of consumer products. We hope that the research results from this study can provide some reference for future researchers to extend uncertainty theory from the perspective of using the specific hazard incidents mentioned in online reviews of consumer products.

In the mid-1960s, there was a growing number of activists, safety experts, and regulatory agencies that came to believe that customers had uncertainty in anticipating the “hidden dangers” of technologically advanced products. At this time, regulators started enforcing regulations on stakeholders, such as manufacturers, who should be held responsible for anticipating the risks associated with the product usage by the consumers (Mohun, 2013). Hence, our study should draw attention to the manufacturers’ need to redesign products with a specific risk level, ranging from low to serious risk, determined by our risk-assessment model. In addition, safety regulators, such as the CPSC, could execute their strategic plan in implementing their listed goals regarding rigorous hazards identification and mitigation. As a result, the CPSC could use our risk-assessment methodology for detecting and prioritizing the hazards of consumer products. By using the risk matrix chart, additional stakeholders, such as insurance underwriters providing product liability insurance, could estimate the insurance-claim likelihood and magnitude, using the likelihood of hazard and severity level for a specific product model. Hence, our risk assessment methodology covers the benefits of a wide range of stakeholders for rapidly identifying and mitigating the specific hazards of consumer products.

We use the assessments of product safety experts for calibrating and validating our computerized product risk-assessment scoring model. While the comparison to expert risk

adjudications revealed that the machine (automated) risk model was severely limited, we can conclude that the model still provided some helpful indicators of product risk. There is clearly a need for substantial future work that continues to use human expert judgment to restructure and recalibrate the risk-assessment model.

References

- Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6), 975-990.
- Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems*, 55(4), 871-882.
- Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, 54(1), 87-97.
- Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International journal of medical informatics*, 100, 108-120.
- Alemzadeh, H., Iyer, R. K., Kalbarczyk, Z., & Raman, J. (2013). Analysis of safety-critical computer failures in medical devices. *IEEE Security & Privacy*, 11(4), 14-26.
- Allen, L.P., Comolli, R., & Heumann, S. (2008). China product recalls: what's at stake and what's next. NERA Economic Consulting.
- Bapuji, H., & Beamish, P. (2008). Product recalls: avoid hazardous design flaws. *Harvard Business Review* 86 (3), 23-26.
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in psychology*, 8, 456.
- Bondy, J., Lipscomb, H., Guarini, K., & Glazner, J. E. (2005). Methods for using narrative text from injury reports to identify factors contributing to construction injury. *American journal of industrial medicine*, 48(5), 373-380.
- Boutros, B. A. S., Hewedi, M. M., Roberts, K. R., & Megahid, F. M. (2014). Food safety traceability systems in the maritime catering logistics. *Food and Nutrition Sciences*, 5(15), 1447.
- Brynjolfsson, E., Hu, Y., & Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49(11), 1580-1596.
- Chen, L., Vallmuur, K., & Nayak, R. (2015). Injury narrative text classification using factorization model. *BMC medical informatics and decision making*, 15(1), S5.

- CPSC. (2019a). CPSC ALERT: CPSC and Fisher-Price Warn Consumers About Fisher-Price Rock 'N Play Due to Reports of Death When Infants Roll Over in the Product. Retrived from <https://www.cpsc.gov/content/cpsc-alert-cpsc-and-fisher-price-warn-consumers-about-fisher-price-rock-%E2%80%98n-play-due-to>
- CPSC. (2019b). Fisher-Price Recalls Rock 'n Play Sleepers Due to Reports of Deaths. Retrieved from: <https://www.cpsc.gov/Recalls/2019/fisher-price-recalls-rock-n-play-sleepers-due-to-reports-of-deaths>
- CPSC. (2009). Maclaren USA recalls to repair strollers following fingertip amputations. Retrieved from <http://www.cpsc.gov/cpscpub/prerel/prhtml10/10033.html>.
- Dhar, V., & Chang, E. A. (2009). Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing*, 23(4), 300-307.
- Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems*, 40(2), 213-233.
- Gallozzi, M., Tucker, S. (2007). Insuring Against Disaster: Coverage for Product Recalls, The Insurance Coverage Law Bulletin, October, 2007.
- Goldberg, D.M. (2019). Automated extraction of product feedback from online reviews: Improving efficiency, value, and total yield (Doctoral dissertation). Virginia Tech.
- Goldberg, D. M., & Abrahams, A. S. (2018). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*, 105, 52-65.
- Hora, M., Bapuji, H., & Roth, A. V. (2011). 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), 766-777.
- Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67, 84-94.
- Lee, J. H., Kim, Y. H., & Kwon, J. H. (2012). Fatal misuse of humidifier disinfectants in Korea: importance of screening risk assessment and implications for management of chemicals in consumer products. *Environmental Science & Technology*, 46(5), 2498-2500.
- Lee, M., Kim, J. H., Lee, D., Kim, J., Lim, H., Seo, J., & Park, Y. K. (2018). Health risk assessment on hazardous ingredients in household deodorizing products. *International journal of environmental research and public health*, 15(4), 744.
- Liu, Y., Feng, J., & Liao, X. (2017). When Online Reviews Meet Sales Volume Information: Is More or Accurate Information Always Better?. *Information Systems Research*, 28(4), 723-743.
- Marucci-Wellman, H., Lehto, M., & Corns, H. (2011). A combined Fuzzy and Naïve Bayesian strategy can be used to assign event codes to injury narratives. *Injury Prevention*, 17(6), 407-414.

- Marucci-Wellman, H. R., Corns, H. L., & Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review. *Accident Analysis & Prevention*, 98, 359-371.
- Mohun, A. (2013). *Risk: negotiating safety in American Society*. JHU Press.
- Mummalaneni, V., Gruss, R., Goldberg, D. M., Ehsani, J. P., & Abrahams, A. S. (2018). Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety science*, 104, 260-268.
- Nanda, G., Grattan, K. M., Chu, M. T., Davis, L. K., & Lehto, M. R. (2016). Bayesian decision support for coding occupational injury data. *Journal of safety research*, 57, 71-82.
- Nasri, L., Baghersad, M., Gruss, R., Marucchi, N. S. W., Abrahams, A. S., & Ehsani, J. P. (2018). An investigation into online videos as a source of safety hazard reports. *Journal of Safety Research*.
- Pan, S., Wang, L., Wang, K., Bi, Z., Shan, S., & Xu, B. (2014). A Knowledge Engineering Framework for Identifying Key Impact Factors from Safety-Related Accident Cases. *Systems Research and Behavioral Science*, 31(3), 383-397.
- Pan, Y., Rose, C. E., Haber, M., Ma, Y., Carrasco, J. L., Stewart, B., ... & McNeil, M. M. (2013). Assessing agreement of repeated binary measurements with an application to the CDC's anthrax vaccine clinical trial. *The international journal of biostatistics*, 9(1), 19-32.
- Peachman, R. (2019). Fisher-Price Rock 'n Play Sleeper Should Be Recalled, Consumer Reports Says. Retrieved from <https://www.consumerreports.org/recalls/fisher-price-rock-n-play-sleeper-should-be-recalled-consumer-reports-says/>
- Saaty, T. L. (2003). Decision-making with the AHP: Why is the principal eigenvector necessary. *European journal of operational research*, 145(1), 85-91.
- Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems. *Journal of Management Information Systems*, 34(4), 1054-1081.
- Soper, H. E., Young, A. W., Cave, B. M., Lee, A., & Pearson, K. (1917). On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and RA Fisher. *Biometrika*, 11(4), 328-413.
- Staelin, R. (1978). The effects of consumer education on consumer product safety behavior. *Journal of Consumer Research*, 5(1), 30-40.

- Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R., & French, D. D. (2009). Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, 10(4), 253.
- Trottman, M., & Mitchell, J. (2010). Toyota recall delay comes under scrutiny. *Wall Street Journal*, January 29, 2010, accessed on: January 30, 2010.
- Union, E. (2016). *EU general risk assessment methodology (Action 5 of Multi-Annual Action Plan for the surveillance of products in the EU)*. Retrieved from: <https://ec.europa.eu/docsroom/documents/17107/attachments/1/translations/en/renditions/pdf>
- Vallmuur, K., Marucci-Wellman, H. R., Taylor, J. A., Lehto, M., Corns, H. L., & Smith, G. S. (2016). Harnessing information from injury narratives in the ‘big data’era: understanding and applying machine learning for injury surveillance. *Injury prevention*, 22(Suppl 1), i34-i42.
- Watson, W. L., & Ozanne-Smith, J. (1995). *Consumer Product-related Injury in Australia: Hospital & Medical Costs to Government*. Monash University Accident Research Centre.
- Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision support systems*, 90, 23-32.
- Wu, Q., Li, Y., Lu, X., & Liu, X. (2017, November). Risk assessment method for physical hazards of consumer products based on uncertainty of use scenarios. In *2017 9th International Conference on Advanced Infocomm Technology (ICAIT)* (pp. 443-449). IEEE.
- Zhang, X., Qiao, Z., Ahuja, A., Fan, W., Fox, E. A., & Reddy, C. K. (2019). Discovering Product Defects and Solutions from Online User Generated Contents.

Appendix A: Survey Instrument (Sample)

Q1a Please indicate the number of years experience you have in the following areas:

0 10 20 30 40 50 60 70 80 90 100

Product Safety ()	
Product Risk Assessment ()	

Q1b Do you hold any certification(s) or degree(s) related to product safety and/or product risk assessment?

☐ Yes (1)

☐ No (2)

Q1c What specific certifications and/or degrees you hold that relate to product safety and/or risk assessment?

Q2 Please indicate your opinion on the severity of the following hazard categories on a scale of 0 (not at all severe) to 1 (extremely severe).

	0	0	0	0	0	1
Chemical hazard ()						
Choking hazard ()						
Cutting hazard ()						
Fire hazard ()						
Hitting hazard ()						
Strangulation hazard ()						
Suffocation hazard ()						
Swallowing hazard ()						

Q4a

Please click the [hyperlink](#) to look through the customer reviews of the product model:
XXXXXXXXXXXXX.

Q4b After reading the consumer reviews, which (if any) of the following hazard types did you observe?

☐ Chemical hazard (1)

☐ Choking hazard (2)

☐ Cutting hazard (3)

☐ Fire hazard (4)

☐ Hitting hazard (5)

☐ Strangulation hazard (6)

☐ Suffocation hazard (7)

☐ Swallowing hazard (8)

☐ Thermal hazard (9)

Q4c

The following risk matrix (below) assesses product safety risk based on the combination of the severity of harm and the probability of harm occurring.

After reading the consumer reviews (above), please choose one circle that corresponds to the combination

of the severity of harm and the probability of harm that you would estimate for the specific product model.

	1 (lowest severity) (1)	2 (2)	3 (3)	4 (highest severity) (4)
> 50% (highest probability) (1)	☐	☐	☐	☐
> 1/10 (2)	☐	☐	☐	☐
> 1/100 (3)	☐	☐	☐	☐
> 1/1,000 (4)	☐	☐	☐	☐
> 1/10,000 (5)	☐	☐	☐	☐
> 1/100,000 (6)	☐	☐	☐	☐
> 1/1,000,000 (7)	☐	☐	☐	☐
< 1/10,000,000 (lowest probability) (8)	☐	☐	☐	☐

Q4d The following sentences show the specific hazard reports that are found about the product model:

I wouldn't recommend to anyone to put a baby in a chemical smelling crib and good warning to you for all new cribs being purchased off gas in a room with a window open for months before your baby sleeps in them.

This product came within a day and I was thrilled...until I pulled parts out of box to discover a terrible chemical fume smell.

Q4e After reading the hazard reports (above), which (if any) of the following hazard types did you observe?

☐ Chemical hazard (1)

☐ Choking hazard (2)

☐ Cutting hazard (3)

☐ Fire hazard (4)

☐ Hitting hazard (5)

☐ Strangulation hazard (6)

☐ Suffocation hazard (7)

☐ Swallowing hazard (8)

☐ Thermal hazard (9)

Q4f The following risk matrix (below) assesses product safety risk based on the combination of the severity of harm and the probability of harm occurring.

After reading the hazard reports (above), please choose one circle that corresponds to the combination of the severity of harm and the probability of harm that you would estimate for the specific product model.

	1 (lowest severity) (1)	2 (2)	3 (3)	4 (highest severity) (4)
> 50% (highest probability) (1)	☐	☐	☐	☐
> 1/10 (2)	☐	☐	☐	☐
> 1/100 (3)	☐	☐	☐	☐
> 1/1,000 (4)	☐	☐	☐	☐
> 1/10,000 (5)	☐	☐	☐	☐
> 1/100,000 (6)	☐	☐	☐	☐
> 1/1,000,000 (7)	☐	☐	☐	☐
< 1/10,000,000 (lowest probability) (8)	☐	☐	☐	☐

Appendix B. Human Experts' Risk Scores and Rationales of various products

Table 19. Human experts' (survey respondent) risk assessments on various products

Product #	Risk scores Before vs After hazard snippet-reading	Likelihood					Severity				
		Respondent (1)	Respondent (2)	Respondent (3)	Respondent (4)	Respondent (5)	Respondent (1)	Respondent (2)	Respondent (3)	Respondent (4)	Respondent (5)
21	Before	0.0000001	0.0000001	0.01	0.00001	0.01	0	0	0	0	0
21	After	0.0000001	0.0000001	0.01	0.00001	0.0001	0	0	0.25	0	0.25
24	Before		0.0000001			0.5		0			0.75
24	After		0.0000001			0.5		0			0.75
22	Before			0.1		0.000001			0		0.25
22	After			0.1		0.000001			0		0.25
11	Before			0.0001		0.000001			0.5		0.5
11	After			0.0001		0.000001			0.5		0.5
7	Before		0.5					0.75			
7	After		0.5					0.75			
3	Before			0.01	0.001	0.001			0.75	0	0
3	After			0.01	0.01	0.01			0.75	0.75	0.75
14	Before		0.0001					0.50			
14	After		0.0001					0.50			
1	Before			0.01		0.01			0.50		0.75
1	After			0.01		0.01			0.50		0.25
9	Before	0.00001					0.25				
9	After	0.00001					0.25				
5	Before	0.000001				0.00001	0				0.25
5	After	0.000001				0.00001	0				0.5
18	Before	0.0000001	0.0000001		0.0000001		0	0		0	
18	After	0.000001	0.000001		0.0001		0.25	0.50		0.25	
25	Before	0.000001		0.001			0		0.25		
25	After	0.00001		0.001			0		0.25		
8	Before	0.000001					0				
8	After	0.000001					0				
10	Before			0.0001					0.50		
10	After			0.0001					0.50		

(#): represents a survey respondent's ID number (anonymized).

Product #	Risk scores	Likelihood					Severity				
		Respondent (1)	Respondent (2)	Respondent (3)	Respondent (4)	Respondent (5)	Respondent (1)	Respondent (2)	Respondent (3)	Respondent (4)	Respondent (5)
28	Before		0.001			0.000001		0.50			0.75
28	After		0.001			0.000001		0.50			0.75
19	Before	0.000001	0.001		0.000001		0.25	0.75		0	
19	After	0.000001	0.001		0.000001		0.25	0.75		0.25	
4	Before	0.0000001					0				
4	After	0.0000001					0				
26	Before			0.1	0.5				0.25	0	
26	After			0.1	0.1				0.25	0	
6	Before	0.001					0.75				
6	After	0.001					0.75				
27	Before		0.5	0.1				0.75	0.5		
27	After		0.5	0.1				0.75	0.5		
23	Before	0.000001				0.0001	0.25				0.25
23	After	0.000001				0.01	0.25				0.5
15	Before				0.000001					0	
15	After				0.000001					0	
16	Before				0.0000001	0.000001				0	0.5
16	After				0.0000001	0.0000001				0	0.75
17	Before	0.5				0.0001	0.75				0.75
17	After	0.5				0.0001	0.75				0.75
12	Before		0.1		0.001			0.75		0	
12	After		0.01		0.001			0.75		0.50	
29	Before	0.000001					0.50				
29	After	0.000001					0.50				
20	Before		0.000001	0.0001	0.0000001			0	0.25	0	
20	After		0.00001	0.0001	0.0000001			0	0.25	0	

Table 20. Hazard snippets discovered by computer and Rationales given by 5 human experts (survey respondents) for each product

Product #	Hazard Snippet	Respondent	Rationale
21	Chemical: My son has NEVER had any sort of diaper rash until July 7th, 2010.- My infant son began using the diapers on my July 5th (afternoon).- On the evening of July 7th, my son's buttocks began turning a light pinkish color. On July 8th, his rash turned dark pink and smooth. By night time, she was screaming from the pain and a couple of the spots in the rash had started bleeding.	1. (Before: None; After: None)	Research has shown that about 25% of all infants experience diaper rash. There is nothing in any of the 1 and 2-star reviews and in this review to indicate a product defect and associated risk of harm.
		2. (Before: None; After: None)	I reviewed this product while a product safety manager at Amazon. It appeared that customers didn't realize that the danger stayed wet and that this wetness could cause severe rashes. There was no proof whatsoever of any kind of actual harm from the new diapers.
		3. (Before: Chemical; After: Chemical)	I only observed 2 reports of a rash that was purported to be caused by contact with the diaper materials out of about 51 reviews. The rashes cleared, one case without any treatment, after the diaper was no longer used.
		4. (Before: Chemical; After: Chemical)	Rash is minor injury, easily avoidable.
		5. (Before: Chemical; After: Chemical)	I don't know the frequency with which this occurs based on one review, but other reviews make me think it's not that high. Severity rating a bit higher because of distress and bleeding caused to the baby.
24	Chemical: I wouldn't recommend to anyone to put a baby in a chemical smelling crib and good warning to you for all new cribs being purchased off gas in a room with a window open for months before your baby sleeps in them. This product came within a day and I was thrilled...until I pulled parts out of box to discover a terrible chemical fume smell.	1.	
		2. (Before: None; After: None)	The customer misunderstood. Before wood items can be shipped to the United States, particularly California, they must be fumigated. They are then loaded directly onto containers and the chemical smell of the fumigation remains. The likelihood of Stork Craft selling toxic paint on cribs is pretty much nil.
		3. (Before: Chemical; After: Chemical)	n/a
		4. Dennis	n/a
22	Chemical: I had to get rid of it and leave the room in order to breathe freely without irritation. Air fresheners are largely unregulated in terms of their ingredients and toxicity. I found the vapors to be incredibly irritating to my sinuses and overwhelming.	5. (Before: Chemical; After: Chemical)	I'm going to assume that the smell is more of an irritant than an actual hazard
		1.	n/a
		2.	n/a
		3. (Before: Chemical; After: Chemical)	There is very limited data. The one report (out of two total reports) that mentions sinus irritation, doesn't clearly indicate any specific hazard. There is a reference to an MSDS for one of the ingredients, but it is unclear what kind of exposure is of concern. For example, is the concern due to physical contact with the ingredient and/or in what concentrations.
		4.	n/a
		5. (Before: Chemical; After: Chemical; Swallowing)	no idea of probability of the hazard but seems like a bit more than an irritant

n/a: Respondent was not randomly assigned to assess the corresponding product.

Product	Hazard Snippet	Respondent	Rationale
11	Chemical: I lost over \$300.00 worth of food, had to live out of a ice chest for a month for a family of 3 suffered headaches because we could not get the toxic smell out of our home and the new fridge we had to pay MORE money for outta our pockets can't even be delivered for another week! Leaving me with a fire risk fridge we cannot put food into or plug in, and gave the entire household migraines for 3 days due to the toxic fumes in the house. 3 weeks now no fridge, tech shows up (a REAL tech) opens EVERYTHING he can unscrew see's burn marks coming out of base of fridge in the rear smells what I smelt.	1.	n/a
		2.	n/a
		3. (Before: Fire; After: Fire)	There is very limited data on this one. There is a single report of a potential fire risk, but no real evidence of fire. Where there is smoke, there is not always fire.
		4. (Before: Fire; After: Fire)	The potential hazard is fire, but with just one report and so little information it is hard to make any conclusions. Fire hazards in a home always rate a high severity; the probability estimate was nothing more than a guess! Needs more investigation.
		5. (Before: None; After: Chemical, Fire)	For this review, not sure how frequent this hazard is based on one review. but pretty sever if the house burns down.
7	Choking: Some of these small pieces (which slip out easily) could pose a choking hazard. Fortunately I was right there to pull the piece out of his mouth. My second issue is that the material is a little too soft as my son has easily bitten and torn right through some of the letters.	1.	n/a
		2. (Before: Chemical; Choking; After: Chemical, Choking)	This is a proven choking hazard and should certainly be recalled. While they label it 3+, the actual behavior is for smaller children to use the toy. The lead claim is iffy, but believable, because this looks like one of these "overnight" brands that litter Amazon, buying noncompliant junk from Alibaba, then reselling it on Amazon to unsuspecting American consumers.
		3.	n/a
		4.	n/a
		5.	n/a
3	Choking: Even if this is not recommended for children under 3 (because of the serious risk of a choking hazard) it is a shame that it falls apart so easily. Fortunately my son is old enough to not put things in his mouth but this is could be a potential choking risk for younger children. I asked his teacher and she said he wasn't pulling at the eye - it just fell out.	1.	n/a
		2.	n/a
		3. (Before: Choking; After: Choking)	There aren't many reviews, but there are numerous reports of pieces falling off easily, including the eyes (which presumably would fit inside a choke tube). Even if the product is intended for children old enough to know they should not eat the parts of the toy, it is safe to assume younger children may also be using this toy, either because caregivers ignored the warning, or because children of the intended age range often have younger siblings or playmates around.
		4. (Before: Choking; After: Choking)	This product would appeal to children under age 3; the age labeling is incorrect. Assuming that the eye fits entirely within a small parts test cylinder, the injury severity is severe if a young child swallowed it. Probability is fairly high that a young child will put pieces in mouth without a parent noticing.
		5.	

n/a: Respondent was not randomly assigned to assess the corresponding product.

Product	Hazard Snippet	Respondent	Rationale
14	Choking: After seeing someone else who posted their child was actually choking and not breathing - and me seeing the danger based on my experience but acting on it before such a thing happened, I am giving this a 1 star and I strongly believe this product should be recalled.	1.	n/a
		2. (Before: Chemical, Choking, Strangulation; After: Choking)	To be highest severity, the actual product has to have been proven to actually cause harm. We found during investigations that often customers would exaggerate or even fabricate details in reviews. Regarding the frequency of the issue, at 360 reviews, this product likely sold at least 30k units. Only a handful of people actually noted issues.
		3.	n/a
		4.	n/a
		5.	n/a
1	Choking: My only complaint is that the little black things that go on the bottom of the feet pop off easily..(when you move it, or pick it up) Just be sure to look for all the black things, make sure they are all attached, b/c they too could be a choking hazard. There is nothing worse than hearing your child choke and you can't do anything about it. Fall: I am afraid the fall down and his hand get stuck and break his hand.	1.	n/a
		2.	n/a
		3. (Before: Choking, Fall, Hitting, Swallowing; After: Choking)	There were many reports of pieces breaking off and children putting them in their mouths, including reports of choking and difficulty breathing. There is also a hazard from it being too easy for a toddler to climb over and fall; there was one report of a child landing on his head after scaling the gate. There was also at least one report of a part suddenly breaking loose and shooting into the air, striking a child.
		4.	n/a
		5. (Before: Choking, Mechanical, Fall ; After: Fall)	given that the product can move around seems pretty likely that a toddler could fall but I think it's less likely that the fall will result in significant injury.
9	Choking: Imagine my surprise when I saw red paint chips all over my baby's little mouth!	1. (Before: Chemical, Mechanical, Fall; After: Chemical)	Conflicting reports noting paint coming off and some consumers specifically stating that they had no paint peeling. Low likelihood and low severity as any amount chemical exposure is likely to be low and route of exposure ingestion.
		2.	n/a
		3.	n/a
		4.	n/a
		5.	n/a

n/a: Respondent was not randomly assigned to assess risk of the corresponding product.

Bold sentences: refer to a different hazard-specific snippet found by the ZAGG risk model and a different rationale given by the expert for a specific model.

Product	Hazard Snippet	Respondent	Rationale
5	Choking: I spit out a very sharp piece of hard black plastic.	1. (Before: Choking, Mechanical; After: Choking)	Few complaints of debris and little likelihood of severe injury should you swallow a piece of plastic or rubber gasket.
		2.	n/a
		3.	n/a
		4.	n/a
		5. (Before: Choking, Mechanical; After: Mechanical; Swallowing)	no idea of the frequency but the second review I read referring to plastic parts inside the blender so made a guess. Not sure it's a choking hazard since part is very small so maybe mostly swallow.
18	Mechanical: My cut was minor compared to hers but I was being as careful as I could be and I still cut myself. Also, if you pour without the cap on the blade comes tumbling out and you have to get out of it's way so it doesn't cut you. She is on her way to get stitches now. Choking: After just using this a few times, I found small piece of plastic that had broken off of the blade housing in my smoothie. UNTIL I made a smoothie yesterday and choked on a piece of plastic	1. (Before: None; After: Mechanical)	The hazard should be obvious (user expects the blades to be sharp) and the majority of the reviews failed to cite this issue as a problem, leading to the conclusion that the likelihood was low.
		2. (Before: None; After: Swallowing)	Very unusual situation, but can certainly be a risk of swallowing a foreign object.
		3. John	n/a
		4. (Before: Swallowing; After: Mechanical)	Appears to be a poor quality product, plastic cracks and parts don't fit well. The blades falling out unexpectedly during cleaning seems like a significant hazard. No great way to gauge the probability, but I would definitely investigate further and consider pursuing a recall.
		5.	n/a
25	Mechanical: My husband almost took a hammer and destroyed the chair after the last spring popped and cut his finger. Fall: About a week later the ball bearings piece broke while my wife was nursing our son, causing them to fall out of the chair. My husband (about 180 lbs, average build) finished building the chair, sat on it for the very first time, and the base broke immediately on the right side.	1. (Before: Fall; After: Mechanical)	First aid type injury with low likelihood of having a body part close to the spring when it failed. Greater risk is the risk of falling when the chair breaks.
		2.	n/a
		3. (Before: Fall; After: Fall)	There are many reports of chair breaking, and a few reports of people falling. The fall is not far, however, and there are no reports of major injuries. Moreover, the chair seems to give warning in most, if not all, instances that it is about to break. So, this seems to cut down on the incidence level.
		4.	n/a
		5.	n/a

n/a: Respondent was not randomly assigned to assess risk of the corresponding product.

Bold sentences: refer to a different hazard-specific snippet found by the ZAGG risk model and a different rationale given by the expert for a specific model.

Product	Hazard Snippet	Respondent	Rationale
8	Mechanical: The bed has sharp wood-like edges and my kid has 2 bruises from it.	1.	n/a
		2. (Before: None; After: None)	I do not see any safety hazard here. The customer most concerned is the customer who appears to be least able to control and soothe their child. It's obviously cheap junk though.
		3.	n/a
		4.	n/a
		5.	n/a
10	Fire: The interior light of certain LG refrigerator models stays on after the doors are closed, leading to spoiled food and melted plastic, and presenting a serious fire hazard. The lights (two 60 watt bulbs) fry the food and melt the plastic.	1.	n/a
		2.	n/a
		3. (Before: Fire, Swallowing, Thermal; After: Fire, Swallowing, Thermal)	There was only one review mentioning a hazard, but it did mention a class action suit. It is unclear if the suit is for reliability or safety issues. Given there is only one review mentioning it, I am going to conclude is likely reliability. The overheating mentioned could lead to someone getting burned (by touching an unexpectedly hot surface) or might possibly lead to a fire. There aren't enough facts to know whether there is enough heat to ignite the fuel. The other possible hazard that seems more likely is the refrigerator doesn't keep things cool enough, which could lead to someone swallowing spoiled food.
		4.	n/a
		5.	n/a
28	Fire: 3 weeks now no fridge, tech shows up (a REAL tech) opens EVERYTHING he can unscrew see's burn marks coming out of base of fridge in the rear smells what I smelt and informs my retailer or it burning. Leaving me with a fire risk fridge we cannot put food into or plug in...	1.	n/a
		2. (Before: Chemical, Fire; After: Chemical, Fire)	The assessment is based on the fact that there was not a live fire. Some designs pass certifications because the device breaks itself, and is therefore safe for use in the home from a fire perspective, but the chemical risk is significant from breathing toxic plastics and metal particulate matter.
		3.	n/a
		4.	n/a
		5. (Before: Fire; After: Fire)	No idea really of the probability based on a very small number of reviews, but the house might catch fire so severe harm possible

n/a: Respondent was not randomly assigned to assess the corresponding product.

Product	Hazard Snippet	Respondent	Rationale
19	Fire: If I hadn't been in the same room when it malfunctioned, I think it could've burned my house down. Just now, one of my two started stinking like electrical burning, and it was turned off!	1.(Before: Fire; After: Fire)	Considering all the reviews, few incidents of overheating. This particular report is speculative ("I think it could've burned my house down") and not helpful in analyzing the risk. Smells of electrical burning is not a good indicator of a potential for fire.
		2. (Before: Fire; After: Fire)	Based on the quantity of reviews, I'm estimating 10k units or so were sold, potentially more. Multiple complaints reference burning smelling, and one references actual flames. This indicates a serious issue that could result in a serious lawsuit.
		3. John	n/a
		4. (Before: Fire; After: Fire)	While a fire in a home rates the highest severity, this product appears to fail safely other than smells. No suggestion of actual fire/ flame. Has a built-in thermo-couple.
		5.	n/a
4	Fire: Used it three times, it smoked and got shards of rubber everywhere.	1.(Before: Chemical, Fire; After: Fire)	Low power electrical motors will smoke when undersized but the risk of any components actually igniting is extremely low. The internal components do not provide a sufficient fuel source.
		2.	n/a
		3.	n/a
		4.	n/a
		5.	n/a
26	Fall: I'm 66 years old.. they seemed to stick solid to the wall but within an hour they both fell to the floor. I grabbed this so-called safety bar which was correctly attached to my shower wall, next thing I knew, the suction cups detached from the wall and I was waving the stupid thing in the air while catching myself from falling. Had her helper not been with her she would have fallen and certainly injured herself seriously.	1.	n/a
		2.	n/a
		3. (Before: Fall; After: Fall)	There were many reports of the handles detaching with little, and often no, force applied. In other words, they frequently were just falling off the wall. This actually lessens the risk, because it seems more likely to happen when someone is not pulling on it . I could not find any actual reports of falls. The severity is still a bit elevated because the population using the product tends to be more susceptible to injuries from falls, namely the elderly. Still, a fall can have a wide range of outcomes, from no injury to death. So, it is not as severe as, say, choking.
		4. (Before: Fall; After: Fall)	Looks like 1 out of 5 comments suggested the bar does not stay attached. as falls are the leading cause of injury/ death to seniors, this is a reasonably serious hazard.
		5.	n/a
6	Fall: If my granddaughter leans forward, the whole thing tips. This rail is flimsy, has no real support and is dangerous.	1. (Before: Fall; After: Fall)	Rated this a relatively high risk due to the fact that the product is targeted at an at-risk population segment. A fall has more severe consequences for the elderly, driving the severity to a higher level.
		2.	n/a
		3.	n/a
		4.	n/a
		5.	n/a

n/a: Respondent was not randomly assigned to assess the corresponding product.

Product	Hazard Snippet	Respondent	Rationale
27	Hitting: I doubt he would have fallen out but he definately would have bumped his head on the chair next to him. Within the last week, one of the lap buckles detached on 3 separate occasions when my 24 month, 21 pound daughter moved or arched her back. Strangulation: For one, the blow up part to make the seat higher fell forward out of the seat and my daughter was stuck sitting with this big inflatable thing between her legs.	1.	n/a
		2. (Before: Fall; After: Fall)	Infants are a particularly vulnerable population, and falling from the height of a chair to the floor can result in lifelong brain damage, as happened to an older family member of mine. This is a major problem, with very dangerous potential effects in the worst case scenario.
		3. (Before: Fall; After: Fall) (Before: Fall; After: Fall)	The are numerous reports of children listing too far to one side, nearly falling out, and wiggling themselves and/or the restraints loose. The nature of the product is such that the users are likely to fall from a few feet to the ground, potentially landing on their heads due to how they would tumble out of the inflatable seat. The product has numerous reports of it failing in such a way that it leads to a child being in a position where (s)he could fall head-first from a several feet on the floor.
		4.	n/a
		5.	n/a
23	Hitting: More importantly, I can't get the straps to stay taut around the chair; they're always loose after my daughter eats a meal in the chair, which allows her to rock back and forth in it and bang her head against the chair back.	1. (Before: Choking, Fall; After: Fall)	loose straps and a child that does not sit still while in the seat puts seat at risk of falling from the chair.
		2.	n/a
		3.	n/a
		4.	n/a
		5. (Before: Choking, Mechanical, Fall; After: Fall)	Seems like it could happen a lot. Hard to base risk probability on one review so taking others into account, can't help it.
15	Strangulation: The straps dig into the neck if it is on securely.	1.	n/a
		2.	n/a
		3.	n/a
		4. (Before: Hitting; After: Hitting)	Need more information. Two reviews. I am skeptical that young children would wear this, and whether it provides adequate protection. Conversely, as it does seem sized for young children it is seems unlikely that a ball could be thrown with high enough velocity by another young child to cause heart-stopping injury.
		5.	n/a

n/a: Respondent was not randomly assigned to assess the corresponding product.

Bold sentences: refer to a different hazard-specific snippet found by the ZAGG risk model and a different rationale given by the expert for a specific product.

Product	Hazard Snippet	Respondent	Rationale
16	Suffocation: The car seat is great but I can't get it to recline enough so that my 5 month old's head won't fall when he goes to sleep. If their head is too far down on it it only serves to make their heads fall forward.	1.	n/a
		2.	n/a
		3.	n/a
		4. (Before: None; After: None)	Poor strap design might cause a parent to dis-continue use, but most users were not discouraged. No report or suggestion of injury. Not a safety hazard.
		5. (Before: Fall; After: Suffocation)	no idea if a head falling forward can really risk any kind of injury, so making a guess
17	Suffocation: He would suffocate--there's that much give to the mattress.	1. (Before: Suffocation; After: Suffocation)	Infants can easily roll into the gaps between crib sides and mattresses and suffocate as they are unable to extract themselves.
		2.	n/a
		3.	n/a
		4.	n/a
		5. (Before: Chemical, Suffocation; After: Suffocation)	Seems like a well documented risk for SIDS.

n/a: Respondent was not randomly assigned to assess the corresponding product.

Product	Hazard Snippet	Respondent	Rationale
12	Swallowing: We found he had some holes in his intestine leaking out bacteria, said Leadbetter of the Seattle Children's Hospital. In a statement to ABC News, the toy's manufacturer points out that Magnetix has passed all federal safety requirements and that the box features prominent small part hazard warnings. The little pieces are not supposed to be loose, and came out of the larger pieces. Sweet said the toy was defective because the tiny magnets easily came loose. Other parents are now reporting close calls with the same problem - including a 5-year-old hospitalized in Seattle. As a result, bacteria were forced into his bloodstream - killing him. The magnets are so strong they're hard to pull apart - but they are supposed to be encased in plastic pieces.	1. (Before: Swallowing; After: Swallowing)	Product is a general use product and not likely to be used by children under 3 years of age. There was nothing to indicate these were strong magnets as if they pieces are difficult to remove from the board while playing, the game becomes less useful. In an environment where there are infants and toddlers, the magnets pose a small risk but keep in mind this is a travel set that is most likely to be used outside of the home environment, making the likelihood low. Should two or more be swallowed there is a low likelihood of the need for medical intervention.
		2. (Before: Choking; After: Choking)	The assessment of frequency is the likelihood that something like this could happen, given the likelihood that many people have older siblings that also have younger siblings in the home. Strong magnets are also a known gastrointestinal hazard.
		3.	n/a
		4. (Before: Swallowing; After: Swallowing)	It appears the magnets are strong and pull of the toy easily. While the product is labeled as not for children under 3, it would be attractive for play by young children. Swallowing multiple magnets results in a death. I'd push for a recall of this product.
		5.	n/a
29	Swallowing: If one happens to fall on the floor and my little guy sees it, he would surely try to put it in his mouth and magnets can be life threatening to a child if swallowed.	1.	n/a
		2.	n/a
		3.	n/a
		4. (Before: Swallowing; After: Swallowing)	This product is not for use by children under age 3, and may not even be appealing to older children who are still mouthing things. The magnets appear to come loose easily, but also appear to have very low magnetic properties; if in the unlikely event several were swallowed, the actual threat may be limited.
		5.	n/a
20	Thermal: The drive shaft heating up for no reason I could see. I'm taking it back today and buying a better model. Both these were extremely hot, though the motor was not hot and did not have that familiar burnt insulation smell common when a motor over heats.	1.	n/a
		2. (Before: None; After: None)	There is no safety risk to this Oster blender. They replaced a lot of the metal parts in the motors years ago with cheaper plastic parts. When the motor is on for too long (longer than 30 seconds usually) on poorly machined parts, the plastic starts to melt.
		3. (Before: Fire; After: Fire)	There are numerous reports of the motor overheating and smoke being emitted, often followed by the product no longer functioning. This seems like what would be expected in a situation where the motor is overloaded and the thermal fuse inside eventually fails. There is only one report I saw of what might have been an actual fire, which should not happen because of the way the motor is thermally protected.
		4. (Before: None; After: None)	Apparent quality problems but does not appear to present a safety hazard.
		5.	n/a

Chapter 3. Facebook hospital reviews: Automated service quality detection and relationships with patient satisfaction

Abstract

This study analyzes the types and prevalence of service quality concerns present in online hospital reviews. We characterize the association between different types of service attributes and overall ratings. We then demonstrate text analytics using key “smoke” terms to detect these service concerns and thereby determine primary rationales for patient opinions rather than merely indicating the direction of opinions as in conventional sentiment analysis. Analyzing a large dataset of hospital reviews posted on Facebook, seven core types of service attributes were identified to classify the reviews along major dimensions. We utilize regression analysis to explore the association between each attribute and overall star ratings. Then, we employ text and sentiment analyses to identify key terms indicative of these attributes. We find that these service attributes are generally associated with non-emotive terms, indicating limited utility of sentiment analysis. Instead, smoke terms were far more effective at identifying core service attributes. This study demonstrates that a subset of service attributes mentioned in online reviews tend to drive users’ overall ratings. Managerially, this information helps to prioritize the areas in greatest need of improvement. Additionally, generating key terms for a particular service attribute aids health care policy makers and providers in rapidly monitoring specific concerns and adjusting policies or resources to better serve patients.

1. Introduction

The advent of social media and e-commerce has created a treasure trove of user-generated content (UGC) published online, including blogs, chats, tweets, podcasts, and ratings. UGC platforms have

the potential to be instrumental for managers and prospective consumers who wish to assess product or service quality. Over time, these platforms can accumulate vast volumes of content, which may contain patterns and information salient to decision-making. Analyzing online review platforms has been especially fruitful because consumers often post targeted feedback concerning the performance of a product or service (Fan & Gordon, 2014). In this paper, we seek to assess how the content of online reviews delineates different types of service attributes in the hospital industry. Furthermore, we seek to understand relationships between these service attributes and patients' overall star ratings of hospitals. Understanding what drives these ratings has important managerial implications for the hospital industry because the winnowing out and prioritizing of key linguistic terms related to service attributes can help healthcare providers to understand perceptions of their services and to target their efforts to improve quality.

Several studies have used online reviews in an attempt to understand patient feedback, but making sense of the large volume of online discussions has proved to be a challenging problem. For instance, as most online posts do not directly describe service quality, it can be difficult to detect the most pressing posts; Hawkins et al. (2015) found that only 10% of hospital-related tweets concerned service quality. Additionally, prior work has often found that these posts are difficult to classify, as Hawkins et al. (2015) found that 77.3% of the tweets to be classified as “general” feedback. In this paper, our goal is to develop tools to shed light on service quality concerns by classifying each service attribute into specific values. We demonstrate and develop these tools by collecting the patient reviews from one of the most active and popular social media outlets, Facebook. As Lagu et al. (2016) comment, Facebook “opens doors” that provide relevant patient feedback (i.e., online comments), which hospitals can use in their efforts to improve quality of care.

This paper makes three key contributions to the literature. First, while many prior papers have analyzed online media for insights on hospital service quality, the particular attributes examined

by these online media vary greatly between studies (Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson, 2013; Hawkins et al., 2015; Wagland, Richardson, Armes, Hankins, Lennan, & Griffiths, 2015). In this paper, we manually assess a sample of online Facebook reviews to create a taxonomy that reconciles the service attributes considered in prior literature; we also show the prevalence of each service attribute in these online Facebook reviews. Second, we examine associations between each service attribute and star ratings, which consumers use as a holistic measure of their perception of service quality (Chevalier & Mayzlin, 2006; Hu, Koh, & Reddy, 2014). In doing so, we uncover which service attributes are most important to predicting these star ratings. Third, we develop actionable text analytic techniques to rapidly detect online reviews referring to each service attribute. Drawing from prior works on defect detection (Abrahams, Jiao, Wang, & Fan, 2012; Abrahams, Fan, Wang, Zhang, & Jiao, 2015), we generate “smoke terms” that are predictive of specific service attributes. We show that these techniques outperform baseline techniques, such as sentiment analysis, in filtering online reviews pertaining to a service attribute of interest.

2. Literature review

2.1. Online Ratings and Patient Experiences

Consumers are exponentially writing reviews regarding quality of care, rating both hospitals and physicians on social media sites (Greaves et al., 2013; Pai & Alathur, 2018). Such ratings are crucial indicators of differentiating across hospitals and healthcare providers may have significant potential impacts on behavior and decisions of patients (Rozenblum & Bates, 2013). Research has demonstrated that online ratings of patient care are positively associated with various conventional aspects of service and quality. For example, Bardach et al. (2012) investigated the correlation between “commercial website ratings of hospitals (Yelp.com)” and “traditional hospital performance

aspects” (“Hospital Consumer Assessment of Healthcare Providers and Systems,” or HCAHPS), finding substantial relationships between the “Yelp star scores” and “HCAHPS overall survey scores”. Prior research has shown that although not all hospitals collect this data (Yang, Lee, Liu, Shih, Chen, Chou, & Hwang, 2018), those that do experience improved quality of care (Griffiths & Leaver, 2018). Greaves et al. (2012) reinforced this finding by showing major correlations between “web-based patient ratings on the National Health Service (NHS) Choices website and traditional paper-based survey aspects of patients’ experiences in hospitals”. In addition to industry stakeholders, consumers also make use of online reviews, as surveys have mentioned “91% of consumers read online reviews” whether of products or services, and 84% depend on online feedback. The consumer treat these online feedback as “personal recommendations” (BrightLocal, 2016). Negative feedback is particularly concerning, as it tends to weigh more heavily in user opinions than positive feedback (Wolf & Muhanna, 2011). As such, online opinion can greatly sway the expectations of potential consumers and affect sales (Thirumalai & Sinha, 2009).

Additionally, the online content generated by patients could be an useful insight for health care providers to update their services (Rastegar-Mojarad, Ye, Wall, Murali, & Lin, 2015). As an example, information regarding experiences of patients in hospitals is an important element in the “value-based purchasing program” proposed by the Center for Medicare and Medicaid Services (CMS). The HCAHPS survey is now required nationally in the US, providing a standardized survey score for measuring patient satisfaction on hospital care (Campbell & Li, 2017). Beginning in 2017, the US government ties Medicare reimbursements to patient responses on HCAHPS, with poor-scoring hospitals not receiving full reimbursements. However, due to chronic low response rates of about 30 percent, there is concern as to how well these datasets represent patient experiences. Other data sources such as Leapfrog and ProPublica record hospital ratings based on patient safety and surgeon complication rates respectively. As a result, patients may experience difficulty in choosing a

hospital due to the challenge of collecting all the information from these multiple data sources (Rosen, 2016). However, contrary to the response rate of HCAHPS and the limitation of the other data sources, the patient reviews posted on social media websites deliver a unique and diverse perspective for patients and health care providers alike to understand patient satisfaction (Rozenblum & Bates, 2013). Although the overall response rates across all platforms are unknown, utilizing a combination of data sources ensures that more patient feedback is considered.

Rozenblum and Bates (2013) emphasize the growing importance of social media feedback/data in regards to patients' experiences and asserted that this data will increment the conventional surveys and hence segment various aspects of quality of care. Additionally, Greaves et al. (2013) described the prospect of utilizing the "cloud of patient experience" on the Internet for identifying the characteristics of healthcare quality. With many possible indicators of quality of care, the authors suggested the value of comparing between conventional aspects of patient experience (e.g., HCAHPS) and information sourced from online media.

2.2. Text and Sentiment Analysis

Text classification refers to a process in which a set of textual documents is organized into categories based on the content of the documents. Text classification has become popular in medical informatics, such as applications to social media networks to design an early warning system for Adverse Drug Reactions (ADRs) (Yang, Kiang, & Shang, 2015). James et al. (2017) used Latent Dirichlet Allocation (LDA), an unsupervised topic modeling methodology, to examine the primary topics of discussion in patient reviews. Several previous studies (López, Detz, Ratanawongsa, & Sarkar, 2012; Lagu et al., 2016; Ranard, Werner, Antanavicius, Schwartz, Smith, Meisel, Asch, Ungar, & Merchant, 2016; Devarakonda, Mehta, Tsou, Liang, Nowacki, & Jelovsek, 2017)

contributed in segmenting categories of quality issues in the health care domain. Doing-Harris et al. (2016) used a topic-modeling approach to extract the most common topics (e.g., “appointment wait,” “empathy,” “friendliness,” etc.) within only negative comments. Hawkins et al. (2015) classify social media discussions of hospital service attributes, such as “waiting time,” “communication,” “food,” “general,” etc. However, as the majority of patient tweets fall into the “general” category (Hawkins et al., 2015), further research is required to build a better understanding of the dimensions of service quality mentioned in online media.

Sentiment analysis is a well-known text classification technique that seeks to quantify the direction and/or magnitude of the emotive content in a body of text. Sentiment analysis has been applied to a wide range of text analytic domains. For instance, Farhadloo et al. (2016) used sentiment analysis with online TripAdvisor reviews to analyze overall customer satisfaction. Likewise, in healthcare, sentiment analysis had been used to help understand the mood of patients in healthcare/hospital settings (Rodrigues, das Dores, Camilo-Junior, & Rosa, 2016). Traditional sentiment analysis alone, however, poses limitations, especially if the body of text contains opinions, categories or domain-related information impertinent to positive/negative emotive valence. For example, Abrahams et al. (2012) assess the effectiveness of sentiment analysis in uncovering vehicle defects discussed in online media, and the authors concluded that sentiment analysis was an insufficient tool when applied on its own. In fact, many of the terms associated with serious defects, such as the term “airbag,” are non-emotive terms that are germane only within the context of the industry of study.

Previous studies in text analytics created list of domain-related terms for specific target classes (Abrahams et al., 2012; Abrahams et al., 2015; Adams, Gruss, & Abrahams, 2017; Goldberg & Abrahams, 2018; Mummalaneni, Gruss, Goldberg, Ehsani, & Abrahams, 2018). These dictionaries are comprised of specific terms are referred to as “smoke terms,” which are significantly more

prevalent in text of interest (e.g., quality concerns). Importantly, as an alternative to sentiment analysis, smoke terms are tuned to the specific linguistic properties of the industry of interest, so they may include both emotive and non-emotive terms. Smoke terms may be single words (unigrams), but researchers have also extended the methodology to 2-word (bigram) and 3-word (trigram) phrases. Using these terms, researchers and practitioners are able to filter out the most pertinent bits of text for categories of interest. Each term is assigned a weight based on a prevalence metric (Fan, Gordon, & Pathak, 2005) such that new (unseen) text may be scored based on a total “smoke score,” where higher scores reflect greater concerns.

2.3. Research Gap

Table 1 summarizes previous research using text analysis of the Internet and social media in various categories of service attributes. For each study, Table 1 shows the data source, number of records (reviews) analyzed, service attributes analyzed, whether service key terms were analyzed, and whether associations with overall star ratings were analyzed. We classified the service attributes using the classification proposed by Hawkins et al. (2015) with the addition of the patient-reported *diagnosis* (Wagland et al., 2015), waiting time (Nottingham, Johnson, & Russell, 2018) and *incident type* (specialty areas) (Hao & Zhang, 2016; Zhang, Deng, Hong, Evans, Ma, & Zhang, 2018) that were not assessed in that study.

Table 1 highlights the research gap that we seek to address in this paper, specifically the application of semi-automated text classification approach to code types of hospital service attributes in online reviews found in official Facebook pages of hospitals. There is great disparity as to the service attributes captured by these different studies, as each study generally predefines service attributes of interest. Therefore, it is somewhat unclear which service attributes are actually most

prevalent in the text of online reviews, and of these attributes, which have the greatest effect on overall satisfaction.

As a first stage analysis, manual coding of several thousand online reviews can shed light on these issues and offer guidance as to the prevalence of these attributes. Past studies (Greaves et al., 2013; Paul, Wallace, & Dredze, 2013; Wallace, Paul, Sarkar, Trikalinos, & Dredze, 2014; Jung, Hur, Jung, & Kim, 2015; Rastegar-Mojarad et al., 2015; Doing-Harris et al., 2016; Hao & Zhang, 2016) have used various machine-learning approaches to capture latent sentiment across service attributes, but they did not investigate their association with the overall ratings. We also seek to address this question to clarify which service attributes drive patient satisfaction. Our study adapts the techniques proposed by prior work (Abrahams et al., 2012; Abrahams et al., 2015) in generating key terms for detecting reviews related to product or service quality. The findings could benefit hospital administrators as well as regulators to identify the most pressing reviews for each service attribute.

Greaves et al. (2013) suggest generating distinctive phrases for each attribute as opposed to single word-level analysis. Our study adds to the literature (Wallace et al., 2014; Hawkins et al., 2015) not only by using different service attribute, but also by disambiguating by finding distinctive multi-word phrases. This novel information could be rapidly extracted, categorized, and utilized by health care providers and regulators to discover service concerns in hospitals.

Study	Data source(s)	Number of records	Service attribute classification																				Service key term(s)				Association with overall star ratings		
			Environment	Communication	Treatment	Diagnosis	Waiting time	Cost (money)	Incident types	Pain management	Food	General (overall)	System issues	Technical competence	Interpersonal manner	Effective care	Safe care	Medical instructions	Staff interaction	Recommendations	Readmission rates	Bedside manner	Service	Dignity and respect	Emotive words	Emotive phrases		Non-emotive words	Non-emotive phrases
Bardach et al. (2012)	Yelp	6,260 patient ratings	X	X						X							X												X
Bardach et al. (2015)	Yelp, Yahoo, and Google	244 reviews	X	X				X		X			X				X												
Brody and Elhadad (2010)	RateMDs	33,654 reviews			X		X	X						X				X	X										
Campbell and Li (2017)	Facebook and HCAHPS	102 hospitals with star ratings	X	X				X									X												X
Doing-Harris et al. (2016)	Press Ganey satisfaction survey	51,234 free-text responses	X	X			X					X												X	X	X	X		
Glover et al. (2015)	Facebook	679 hospitals																		X									X

		with star ratings																											
Greaves et al. (2012)	NHS Choices site	10,274 posts	X															X											X
Greaves et al. (2013)	NHS Choices site	6,412 posts	X								X											X	X	X	X	X			
Greaves et al. (2014)	Twitter	1,000 tweets	X				X								X	X		X											
Hao and Zhang (2016)	Good Doctor site	731,264 reviews			X				X			X		X						X				X	X	X	X		
Hawkins et al. (2015)	Twitter	11,602 tweets	X	X	X		X	X		X	X	X					X												
Jung et al. (2015)	Local online groups for parents and caregivers	9,450 messages	X	X	X			X					X									X			X	X	X	X	
Lagu et al. (2016)	Facebook	47 comments	X	X			X							X															
Lopez et al. (2012)	RateMDs and Yelp	712 reviews												X	X	X													
Paul et al. (2013)	RateMDs	50,000 reviews												X	X	X									X		X		
Ranard et al. (2016)	Yelp	16,862 reviews		X	X		X	X																					X
Rastegar-Mojarad et al. (2015)	Yelp	6,914 reviews								X															X		X		

Wagland et al. (2015)	National colorectal PROM survey	400 free-text comments		X	X	X												X											
Wallace et al. (2014)	RateMDs	60,000 reviews											X	X	X											X		X	
Our work	Facebook	6,581 reviews	X	X	X	X	X	X	X																X	X	X	X	X

Table 1: Survey of prior work.

3. Research Questions and Contributions

We focused on addressing the following three primary research questions. First (RQ1), which service attributes are present in online reviews of hospitals on Facebook, and how prevalent is each service attribute? Second (RQ2), which service attribute values most strongly influence the final star rating chosen by the patient? Third (RQ3), to what extent can text analytic methodologies be utilized to automatically extract reviews of interest to a particular service attribute?

The first major contribution of this study is in assessing the specific service attribute present in online hospital reviews as well as the relative prevalence of each service attribute. While several papers have examined service attributes for different intentions, Table 1 illustrates that there is a substantial level of variability in researchers' choices of service attributes for analysis. Thus, a holistic approach of coding thousands of reviews for each attribute can unify this literature and clarify the utility of these reviews.

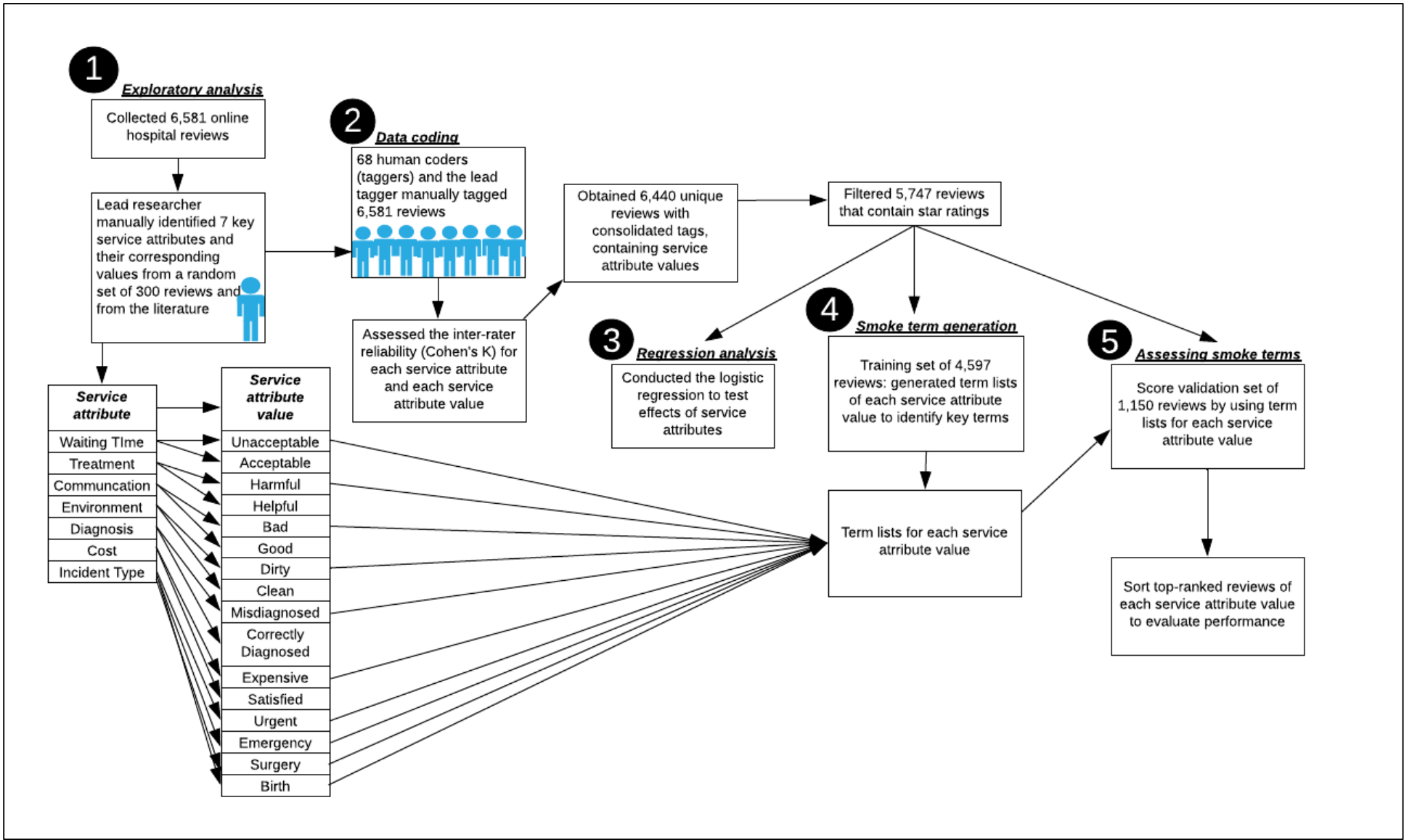
Second, we extend the specific service attributes of the past studies (Hawkins et al., 2015) and test their influence on star ratings, which are generally indicative of consumers' perceptions of quality (Chevalier & Mayzlin, 2006; Hu et al., 2014). Significant relationships have also been observed between patients' star ratings of hospitals and some key terms, specifically key terms referencing mortality and infection rates (Greaves et al., 2012). Investigating patients' complaints and service quality concerns provides an opportunity to evaluate properties of overall patient satisfaction (Johnson & Russell, 2015; Johnson, Russell, & White, 2016). In the United States, 95% of hospitals have a Facebook page, each of which can collect thousands of online reviews, and the overall quantitative ratings provide new insight about patient experience (Rozenblum, Greaves, & Bates, 2017). Understanding the service attributes most associated with these ratings can better explain the mechanisms associated with perceptions of service quality.

Due to the large volume of online reviews on Facebook, it would be cumbersome to manually examine all reviews on hospitals' Facebook pages. Based on previous research related to topic classification for measuring quality of care in U.S. hospitals (Hawkins et al., 2015), our third major contribution is proposing a toolset for investigating different service attributes present in patient reviews of hospitals on Facebook. We seek to identify whether and the extent to which emotive valence (i.e. negative, positive, or neutral) plays a significant role in influencing hospital star ratings compared to the specific terms explicitly mentioned in the reviews. We aim to extract the most prevalent key terms for each service attribute value, which aid in rapid discovery and investigation of specific concerns of hospital services from unstructured online text.

4. Methodology

In this section, we describe the steps involved in the methodology employed in this study. In 4.1, we performed an exploratory analysis of our online reviews to discern which service attributes were most prevalent. Then, in 4.2, we coded a dataset of online reviews based on the service attributes detected. Next, in 4.3, we performed a regression analysis to determine the association between these service attributes and overall star ratings. In 4.4, we used text analytics methods to generate key “smoke” terms capable of rapidly filtering reviews according to these service attributes. Finally, in 4.5, we assessed the performance of these terms. Figure 1 displays a graphic overview of the methodology performed in this study.

Figure 1: Overview of methodology.



4.1. Exploratory Analysis

We chose to analyze user reviews from the top 100 US hospitals identified by Becker's Hospital Review, 2016, reported on Facebook. Thus, these hospitals are widely utilized and ensure thorough geographic coverage of the United States. The top 100 hospitals together accounted for 106,541 beds in 2016. We collected all online hospital reviews available on the Facebook pages of these hospitals. In total, we collected 6,581 online hospital reviews dating from September 24, 2011 to March 09, 2016. Next, the lead researcher manually examined a set of 300 random hospital reviews to determine appropriate service attributes based on these reviews and using the groundwork from the previous studies (see Table 1). This process was initially structured such that the lead researcher searched for service attributes mentioned in prior work. It is important to note that the service attributes mentioned in prior work are not necessarily mutually exclusive. For example, the service attribute "dignity and respect" is very narrow, whereas "communication" is rather broad and likely encompasses dignity, respect, and other factors. Thus, the lead researcher was first tasked with identifying all service attributes mentioned; if substantial overlap existed between service attributes, or if some broad attributes contained narrower sub-attributes, then these issues could be reconciled in ensuing analyses. This analysis identified seven key service attributes to classify the online reviews along core dimensions. These were waiting time, incident type, treatment outcome, communication, hospital environment, diagnosis accuracy, and cost. Each of these attributes is acknowledged in the literature (Dobrzykowski, Callaway, & Vonderembse, 2015; Hawkins et al., 2015; Wagland et al., 2015; Hao & Zhang, 2016; Zhang et al., 2018). Communication was the only service attribute that contained sub-attributes, as we also observed instances of "interpersonal manner," "staff interaction," "bedside manner," and "dignity and respect." As these service attributes overlapped so substantially, we aggregated them into the "communication" service attribute. Interestingly, the analysis did not identify any prevalence of service attributes beyond this seven. For example, while analysis of pain

management has been prevalent in the literature (Bardach et al., 2012; Bardach et al., 2015; Hawkins et al., 2015), the lead researcher did not observe discussions of pain management in our dataset.

4.2. Data Coding

We randomly distributed the reviews to 68 coders, who were undergraduate students studying business at a major public university in the US. We provided the coders with a detailed tagging protocol (see Appendix A) describing how they should label each review. 61 of 68 coders tagged between 100 and 110 reviews; the remaining 7 coders tagged ten or fewer reviews. Overall, we obtained 6,440 unique tags (i.e., 97.9% of total collected reviews). The lead researcher coded a further 300 random reviews for comparison to the students' tags (Abrahams et al., 2015; Goldberg & Abrahams, 2018). For each review, the coder determined whether a specific service attribute value was mentioned in the hospital review. If no code value was identified, the review was tagged as “not mentioned.” We used Cohen's kappa (κ) to determine inter-rater reliability for each code value of each service attribute (specific values are detailed in Appendix B). The procedure we followed to calculate κ was to compare the judgments of Coder A versus Coder B as follows.

- Coder A was designated as the authority coder (the lead member of the research team).
- Coder B was the conservative combined opinion of all other coders. Though voting is often used to determine a final decision in the case of conflicting opinions, the cost of a false negative for concerns of service quality is high. Thus, we employed a conservative strategy, judging a service concern to exist whenever any coder indicated that a concern relating to that service attribute was mentioned (Goldberg & Abrahams, 2018).

We found that κ was consistently greater than 0.4, indicating “moderate” or better agreement (Landis & Koch, 1977) or “fair to good” or better agreement (Fleiss, 1971) between coders.

4.3. Regression Analysis

From the 6,440 reviews tagged in the previous step, we filtered our dataset to retain the 5,747 reviews (87.6% of total collected reviews) that contain a star rating for further analysis. In our study, we found that 65.6% of reviews gave 5 stars; 5.4% gave 4 stars; 2.7% gave 3 stars; 3.6% gave two stars; and 22.7% gave 1 star. This seemingly bimodal distribution is consistent with prior works, which have observed that users with extreme experiences are most likely to post about those experiences online (Hu, Pavlou, & Zhang, 2017). We constructed a model to test whether and the extent to which each service attribute value was associated with the online star rating of hospitals. We used logistic regression to test the extent to which each service attribute value was associated with high or low star ratings.

Each independent variable contains multiple values or levels, e.g., “treatment” has three values: “helpful,” “harmful,” and “not mentioned.” Star ratings, our indicator of perceptions of service quality, were the dependent variable in our model. However, prior works have noted the difficulties in predicting star ratings as a dependent variable, namely the potential for predicting this multinomial dependent variable to yield an overfitted and poorly predictive model (Fürnkranz, 2002; Galar, Fernández, Barrenechea, & Herrera, 2014; King, 2015). Thus, the dependent variable, star ratings, was coded in binary values: ratings from 1 to 3 were coded as “low” (29.0% of reviews), whereas ratings from 4 to 5 were coded as “high” (71.0% of reviews). Thus, this modeling technique enables us to utilize logistic regression and mitigate this potential problem. The model was run to demonstrate the association of each service attribute value with star ratings. This process was performed to ascertain the relative improvement or deterioration in star ratings with increasing mentions of service attributes.

We tested whether one or more of the service attribute values had a strong influence on the online star ratings. We conducted a multi-collinearity test on the variables, which demonstrated that the variables had very low correlations with one another (coefficient less than 0.2), thus confirming the adequacy of internal consistency reliability for all independent variables. In addition to the service attributes, we also utilized the negative word list in a popular conventional sentiment analysis dictionary – AFINN (Nielsen, 2011) – to test the effect of negative words on the star ratings.

4.4. Smoke Term Generation

To identify the key words and phrases associated with each service attribute, we created a *training* set composed of a random 80% of the reviews (4,597) from the full set of 5,747 reviews. We computed the relative prevalence of each term (word or phrase) for each service attribute value using the correlation coefficient (CC) score (Fan et al., 2005) as the specific metric. This technique allowed us to generate lists of single words (unigrams) and two and three word phrases (bigrams and trigrams) specific to each service attribute using the process described in previous studies (Abrahams et al., 2012; Abrahams et al., 2015). We manually curated lists of distinctive or “smoke” terms for each service attribute value to improve generality; consistent with the prior studies (Abrahams et al., 2012; Winkler, Abrahams, Gruss, & Ehsani, 2016), we removed common English words (e.g., “a,” “in,” “for,” “but,” “with,” “they,” etc.). Appendix C contains the top 30 manually curated terms most prevalent in each service attribute.

4.5. Assessing Smoke Terms

We retained a *validation* set containing the remaining 20% of the random 5,747 reviews (1,150). This set of random reviews allowed us to conduct an objective assessment of the performance of automated discovery of each service attribute. We “scored” the validation set by using the term lists of each service attribute value; total “smoke scores” were computed as the sum of the frequency of each smoke term in a review multiplied by that term’s prevalence score (Goldberg & Abrahams, 2018). After computing these scores for each review, we ranked all of the reviews in the validation set from highest score to lowest score, where higher scores indicate greater likelihoods of poor service quality. In order to assess the performance of each smoke scoring method, we computed the area under curve (AUC) statistic. Values of AUC that are close to 1 specify the best models, rarely containing false positives in the top-scored items and rarely containing false negatives in the bottom-scored items.

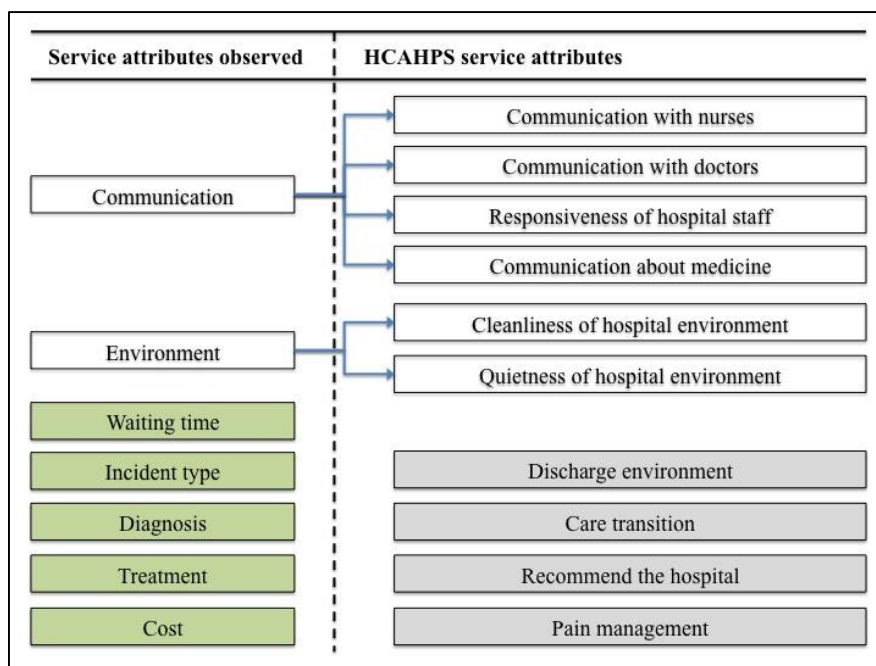
5. Results and Discussion

5.1. Exploratory Analysis

Addressing RQ1, our initial screening of 300 random Facebook hospital reviews revealed seven major service attributes: communication (good, bad, or not mentioned); environment (clean/sanitary, dirty/unsanitary, or not mentioned); waiting time (acceptable, unacceptable, or not mentioned); incident type (urgent, emergency, surgery, birth, or not mentioned); diagnosis (correct, misdiagnosed, undetermined, or not mentioned); treatment (helpful, harmful, or not mentioned); and cost (too expensive or satisfactory/not mentioned). These service attributes overlap somewhat with other studies (see Table 1), although this specific combination of service attributes appears unique to our

study. Interestingly, our service attributes partially map to the HCAHPS survey, a conventional intelligence-gathering tool (see Figure 2). We observed communication and environment as two of our service categories, and these categories overlap with a range of HCAHPS service attributes. The HCAHPS survey does not cover waiting time³³, incident type diagnosis, treatment, or cost. This study did not cover a range of the survey items (shaded in grey) such as discharge information, care transition, recommend the hospital, and pain management.

Figure 2: Mapping of service attributes to HCAHPS.



5.2. Tagging Analysis

Next, our team of 68 undergraduate students assessed a large dataset of 5,747 reviews using the tagging protocol described in Appendix A. Furthering our analysis of RQ1, this step allows us to

³³ Although HCAHPS does not consider waiting time, many hospitals track emergency room waiting time as a component of measures such as NEDOCS (National Emergency Department Overcrowding Scale) [47].

assess the prevalence of each service attribute value in our dataset; these statistics are reported in Table 2.

Service attribute	Service attribute value	Prevalence (percentage of all reviews)		
		Low star ratings	High star ratings	All star ratings
Waiting time	Acceptable	0.28%	2.82%	3.10%
Waiting time	Unacceptable	7.81%	1.17%	8.98%
Waiting time	Not mentioned	20.95%	66.97%	87.92%
Incident type	Urgent	2.61%	3.90%	6.51%
Incident type	Emergency	4.87%	5.19%	10.06%
Incident type	Surgery	2.12%	8.44%	10.56%
Incident type	Birth	0.96%	4.82%	5.78%
Incident type	Not mentioned	18.48%	48.62%	67.10%
Diagnosis	Correct	1.44%	12.22%	13.66%
Diagnosis	Misdiagnosed	2.51%	0.30%	2.80%
Diagnosis	Undetermined	1.18%	0.42%	1.60%
Diagnosis	Not mentioned	23.91%	58.03%	81.94%
Treatment	Helpful	1.74%	30.36%	32.10%
Treatment	Harmful	6.37%	0.38%	6.75%
Treatment	Not mentioned	20.93%	40.21%	61.14%
Communication	Good	1.84%	41.22%	43.07%
Communication	Bad	15.97%	1.18%	17.16%
Communication	Not mentioned	11.22%	28.55%	39.78%
Environment	Clean/sanitary	0.44%	4.18%	4.61%
Environment	Dirty/unsanitary	2.02%	0.54%	2.56%
Environment	Not mentioned	26.59%	66.24%	92.83%
Cost	Too expensive	2.30%	0.78%	3.08%
Cost	Not mentioned/satisfactory	26.74%	70.18%	96.92%

Table 2: Prevalence of service attribute values.

Our service attributes were mentioned very frequently in these reviews. The reviews mentioned 1.72 service attributes on average. 77.92% of reviews mentioned at least one service attribute, and 50.27% mentioned multiple service attributes. For the purposes of our analysis, we coded 1-3 star reviews as “low” and 4-5 star reviews as “high.” Most of the relationships between service attribute values and star ratings are intuitive, but several of the effects are substantial and noteworthy. For example, waiting time was noteworthy, as patients mentioning acceptable waiting times were more than 10 times more likely to post a high-star review, and patients mentioning unacceptable waiting times were more than 6 times more likely to post a low-star review. Additionally, regardless of the type of incident that the patient reported, patients that specifically mentioned some incident were more likely to post a high-star review. Generally, these patients may have come to the hospital with serious health concerns that they were glad to resolve. Communication seemed to be of particular importance to patients, as 60.23% of reviews mentioned communication. Patients that mentioned good communication were over 22 times more likely to post a high-star review, while patients that mentioned bad communication were over 13 times more likely to post a low-star review.

We found that traditional sentiment analysis (here measured using the AFINN negative sentiment dictionary (Nielsen, 2011)) was strongly predictive of overall star ratings, yielding a correlation of 0.54. As a further analysis of star ratings, we also performed an ANOVA to examine the extent to which reviews of certain star ratings differed via pairwise comparisons. A Tukey-Kramer test revealed that the average AFINN sentiment score for reviews of a given star rating differed from the average AFINN sentiment score for reviews of all other star ratings. We display this analysis in Table 3 below. As expected, 5-star reviews were associated with the most positive sentiment on average, followed by 4-star, 3-star, 2-star, and 1-star reviews.

Star rating	Connecting letters report	Mean AFINN sentiment score
5	A	6.35
4	B	4.46
3	C	2.60
2	D	-0.04
1	E	-2.30

Table 3: Toker-Kramer analysis of AFINN sentiment scores at each star rating.

5.3. Regression Analysis

Addressing RQ2, we performed a multiple logistic regression analysis to examine relationships between independent variables (i.e., service attribute values and sentiment values) and overall star ratings (high or low). We display the logistic regression equation below in (1), where P is the probability of a “high” star rating. Most of the variables were binary variables derived from our tagging of service attributes, in which case the variable was equal to 1 if that service concern was indicated and 0 otherwise. Two exceptions are the intercept, β_0 , and the negative sentiment score assessed using the AFINN dictionary.

$$\begin{aligned} \ln\left(\frac{P}{1-P}\right) = & \beta_0 + \beta_1 \text{waiting time unacceptable} + \beta_2 \text{waiting time acceptable} + \\ & \beta_3 \text{incident type birth} + \beta_4 \text{incident type emergency} + \beta_5 \text{incident type surgery} + \\ & \beta_6 \text{incident type urgent} + \beta_7 \text{treatment helpful} + \beta_8 \text{treatment harmful} + \\ & \beta_9 \text{communication good} + \beta_{10} \text{communication bad} + \beta_{11} \text{environment dirty/unsanitary} + \\ & \beta_{12} \text{environment clean/sanitary} + \beta_{13} \text{diagnosis correctly diagnosed} + \\ & \beta_{14} \text{diagnosis incorrectly diagnosed} + \beta_{15} \text{diagnosis undetermined} + \\ & \beta_{16} \text{cost too expensive} + \beta_{17} \text{AFINN negative} \end{aligned} \quad (1)$$

The results of this regression analysis are displayed in Table 4, and 95% confidence intervals for each variable are shown in Figure 3. Variables whose regression coefficients significantly differed from zero at the 0.05 level are **bolded** in Table 4.

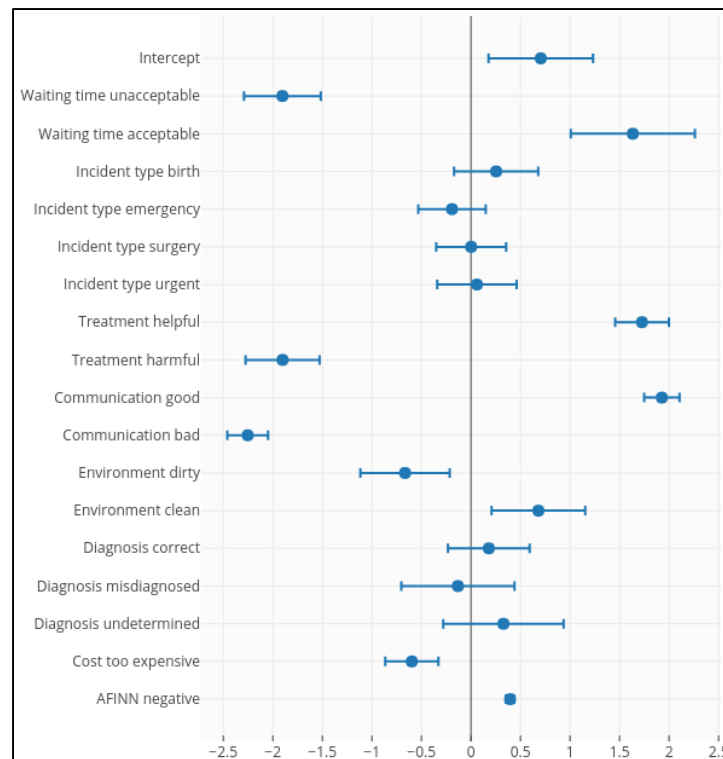
Variable	β coefficient	e^{β}	Standard error	X^2 value	p-value
Intercept	0.71	2.02	0.27	6.85	<0.01
Waiting time unacceptable	-1.90	0.15	0.20	92.54	<0.01
Waiting time acceptable	1.63	5.10	0.32	26.15	<0.01
Incident type birth	0.25	1.29	0.22	1.38	0.24
Incident type emergency	-0.19	0.83	0.17	1.21	0.27
Incident type surgery	0.00	1.00	0.18	0.00	0.99
Incident type urgent	0.06	1.06	0.21	0.09	0.77
Treatment helpful	1.73	5.64	0.14	155.88	<0.01
Treatment harmful	-1.90	0.15	0.19	99.50	<0.01
Communication good	1.93	6.87	0.09	441.24	<0.01
Communication bad	-2.25	0.11	0.11	457.27	<0.01
Environment dirty/unsanitary	-0.67	0.51	0.23	8.37	<0.01
Environment clean/sanitary	0.68	1.98	0.24	7.96	<0.01
Diagnosis correctly diagnosed	0.18	1.20	0.73	0.39	0.39
Diagnosis misdiagnosed	-0.13	0.88	0.29	0.20	0.65
Diagnosis undetermined	0.33	1.39	0.31	1.12	0.29
Cost too expensive	-0.60	0.55	0.14	18.99	<0.01
AFINN negative	-0.40	0.67	0.02	321.19	<0.01

Table 4: Logistic regression analysis output.

Our model had an R^2 value of 0.58, indicating that the model explained 58% of the variance in the star ratings. We also assessed model fit using AUC (area under the curve), where values closer to 1 indicate better fitting models, and values closer to 0 indicate worse fitting models. We observed an AUC value of 0.95, indicating an extremely high degree of fit³⁴.

³⁴ All Variance Inflation Factor (VIF) scores were below 5, indicating that multicollinearity was not a substantial concern for our model. In fact, all VIF scores were below 3 except for “diagnosis undetermined” (3.38).

Figure 3: 95% confidence intervals for regression coefficients.



The logistic regression coefficients provide odds ratios, i.e., the probability of receiving high star ratings (p) divided by the probability of receiving low star ratings ($1 - p$). For instance, a hospital review with the waiting time indicated as acceptable suggests that the review is 5.1 times as likely to be associated with a high star rating (i.e., $e^{1.63}$). Even when controlling for the other service attributes, communication again seemed to be the variable with the greatest effect on overall star ratings, as patients mentioning good communication were over 6 times as likely to give the hospital a high star rating. Mentions of helpful treatments and clean environments were also highly predictive of high star ratings. Interestingly, incident types and diagnoses did not seem to be strongly associated with star ratings, as all of the values for these service attributes were statistically insignificant. Our study adds to that of the past study by Hawkins et al. (Hawkins et al., 2015) by demonstrating that specific

service attributes that were mentioned in hospital-related Facebook posts had significant value in predicting the users' star ratings. Our study demonstrates that the text of online reviews provides critical content for uncovering the service attributes that most influence star ratings.

5.4. Text Analytic Analysis

As a final step in our analysis and to address RQ3, we generated specific terms predictive of each service attribute value. In doing so, we provide a means for hospital administrators to rapidly filter through online feedback and prioritize those reviews pertaining to particular types of feedback. As past studies (Greaves et al., 2013; Paul et al., 2013; Wallace et al., 2014; Jung et al., 2015; Rastegar-Mojarad et al., 2015; Doing-Harris et al., 2016; Hao & Zhang, 2016) mostly focused on single words, or unigrams, we examined the extent to which our terms overlapped with these prior studies. The top 30 most prevalent terms for each service attribute value had very little overlap with each other across each of the service attributes (see Appendix C). Additionally, the most prevalent terms for each service attribute value overlapping with those from prior studies were generally sentiment words. Contrary to the past studies, our study discovered more non-emotive than emotive terms for each service attribute value. Additionally, generating multi-word phrases reveals more details as to the content of specific reviews. For example, we found trigrams such as “for a bed” and “in the hallway” corresponding to unacceptable waiting times, whereas prior studies found only unigrams such as “hours,” “waiting,” “wait,” “hour,” etc. These trigrams provide specificity to forms of patient complaints, such as what the patients were waiting for (waiting to see a doctor versus waiting for a hospital bed) and where they were waiting (in a waiting room, reception area, or hallway).

We also used the validation set to perform an assessment of the performance of these terms in rapidly filtering reviews pertaining to particular service attribute values. Using the “smoke scores”

generated for each review in the validation set, we ranked the reviews in the validation set from most likely (highest score) to least likely (lowest score) to refer to each service attribute value. In Figure 4, we provide lift charts showing the performance of the hospital-specific terms as opposed to generic sentiment analysis. These lift charts show each method’s performance at detecting true positive reviews for each service attribute value (i.e., the number of true positives detected in the top N -ranked reviews). Unigrams, bigrams, and trigrams are compared to the AFINN negative sentiment dictionary and a random chance baseline. Methods showing quicker vertical ascent on the y-axis are better performing. We also show Area Under the Curve (AUC) values in Table 5, which confirm the utility of these methods relative to the AFINN negative sentiment dictionary. In Table 6, we explore the manner in which performance varies as a function of the sample size. Thus, we performed an analysis on smoke term generation for “waiting time unacceptable” in which we varied the training set size. In each trial, we created a balanced training set in which half of the reviews pertained to unacceptable waiting times. We retained the top 200 unigrams (removing stop words), bigrams, and trigrams for evaluation purposes, although we also note the growth in the number of candidate smoke terms at each training set size. Our results indicate that the text analytic techniques perform well even at small sample sizes. However, tagging a greater number of reviews has potential to improve performance further as the size and coverage of the lexicon increases, particularly for longer n -grams. In our study, our coders contributed a total of 69.87 person-hours and generated 6,791 tags. They worked at a rate of 97.20 reviews per hour on average with a standard deviation of 32.79 reviews per hour.

Service attribute	Service attribute value	Unigrams	Bigrams	Trigrams	AFINN negative
Waiting time	Unacceptable	0.87	0.83	0.72	0.75
Incident type	Emergency	0.84	0.47	0.71	0.62
Incident type	Birth	0.92	0.84	0.79	0.60

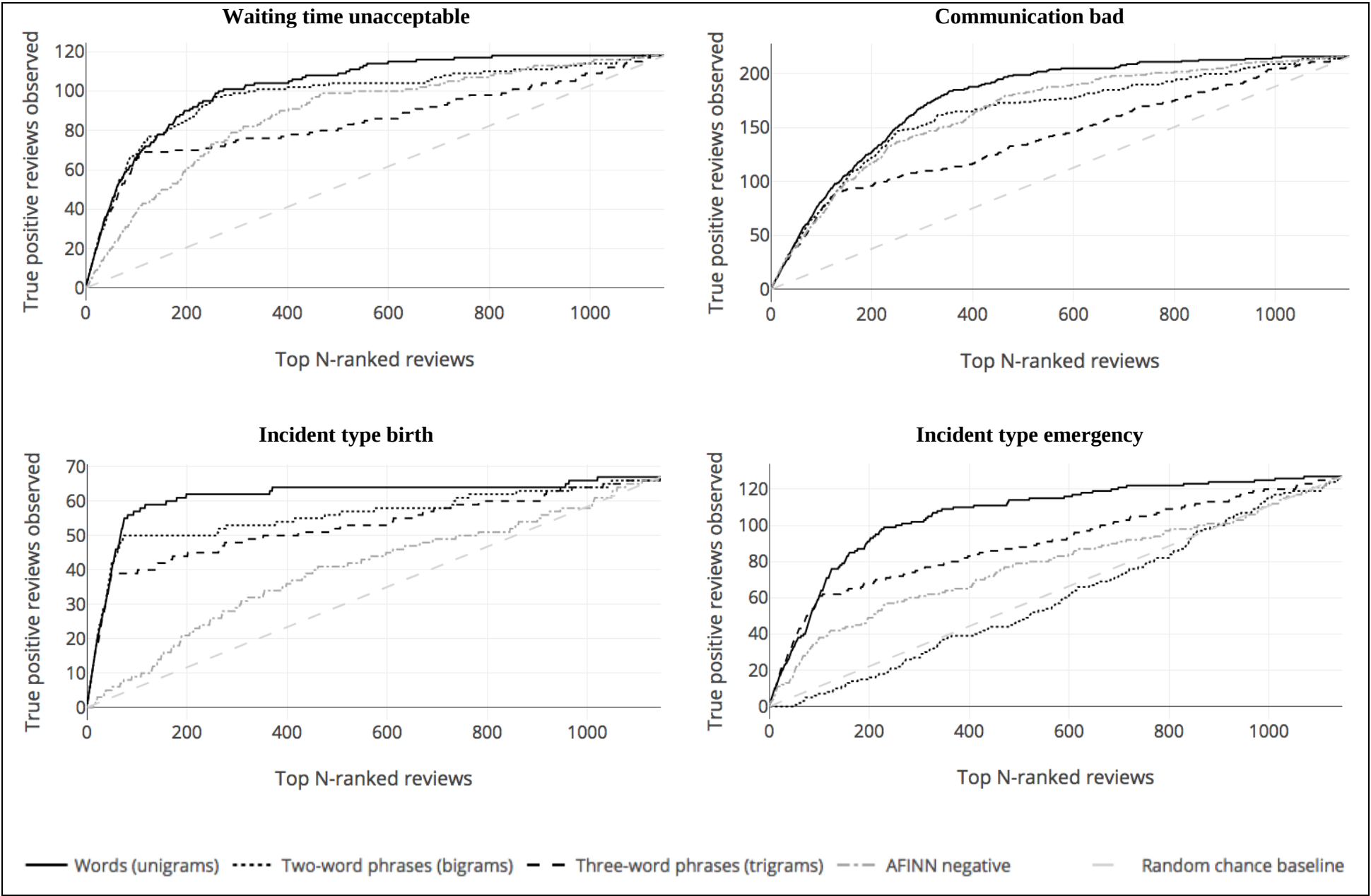
Incident type	Surgery	0.86	0.78	0.51	0.62
Incident type	Urgent	0.78	0.68	0.57	0.57
Treatment	Harmful	0.87	0.74	0.63	0.81
Communication	Bad	0.82	0.77	0.66	0.77
Environment	Dirty	0.77	0.64	0.48	0.69
Diagnosis	Misdiagnosed	0.94	0.82	0.71	0.85
Cost	Too expensive	0.90	0.82	0.78	0.76

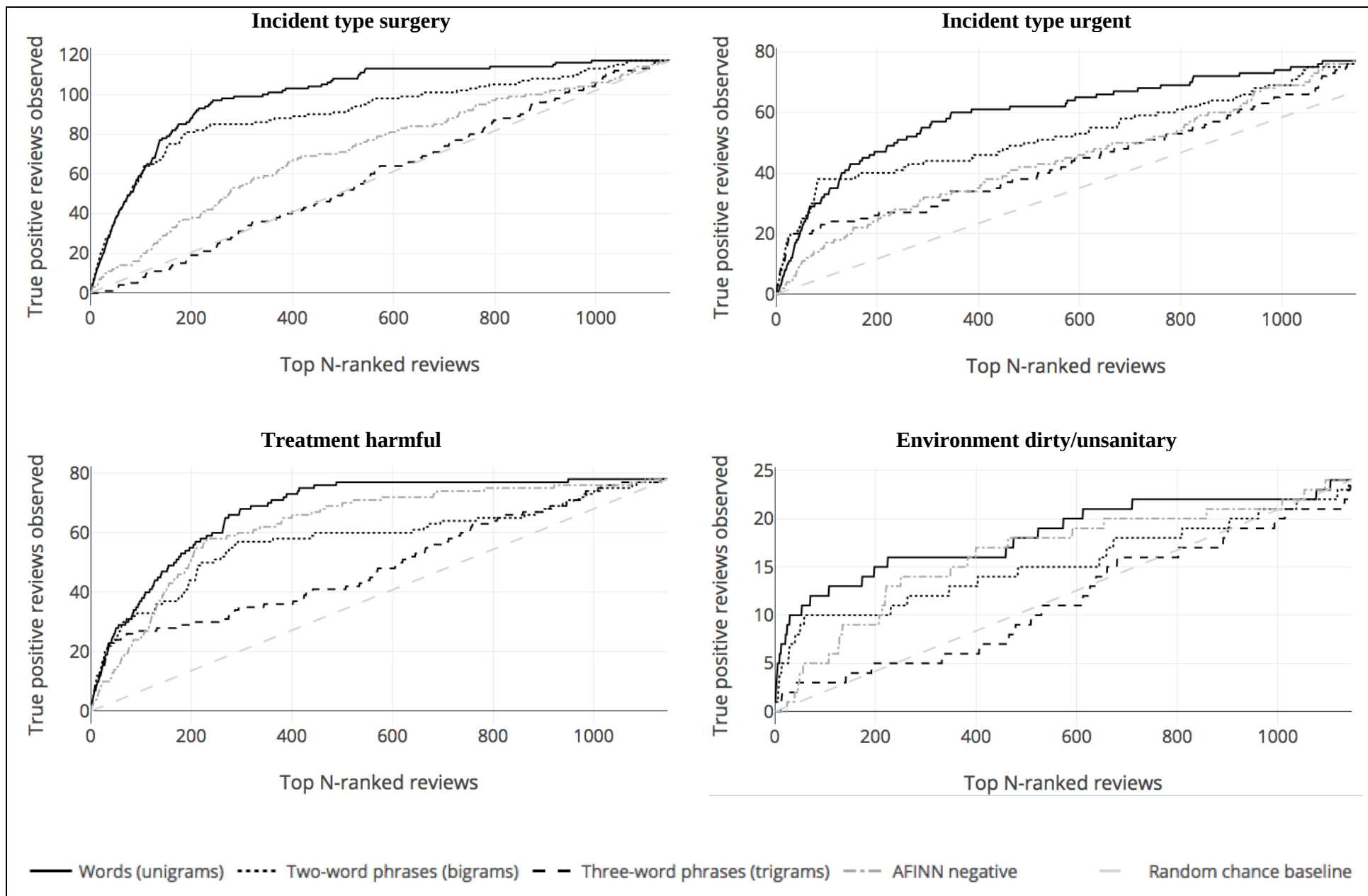
Table 5: Comparison of Area Under the Curve (AUC) values across methods and service attribute values.

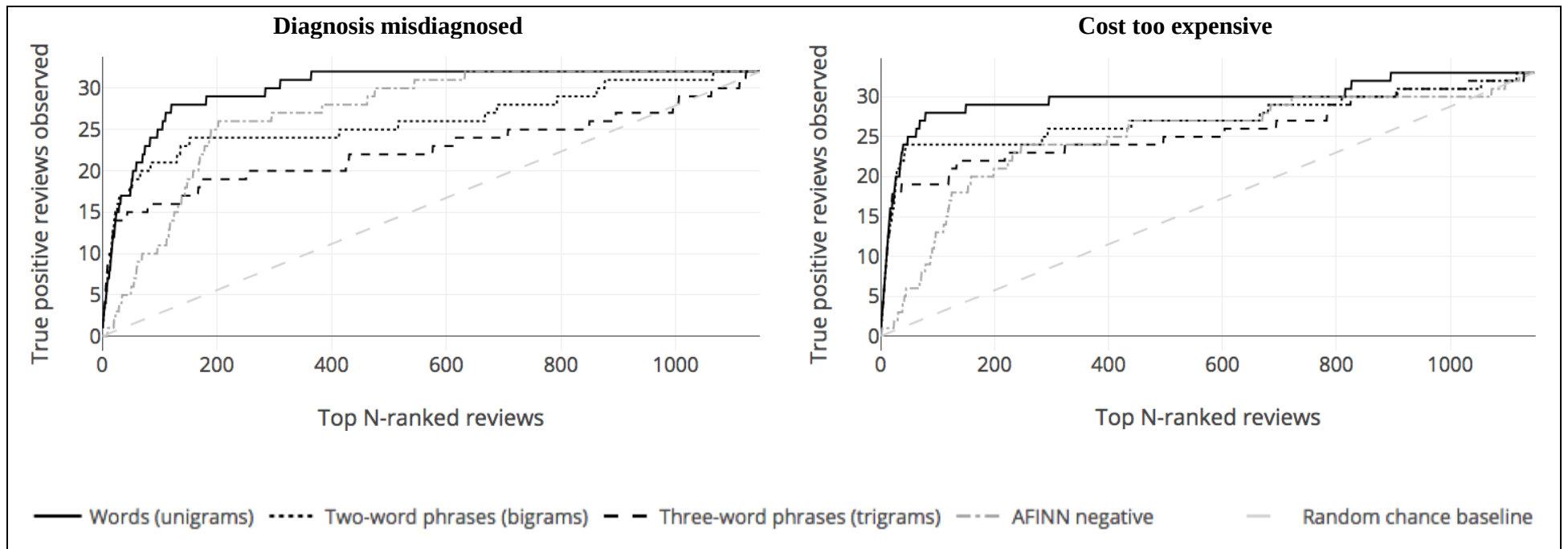
Training set size (count of positives)	AUC (count of unique candidate terms in training set)		
	Unigrams	Bigrams	Trigrams
100 (50)	0.84 (1,285)	0.79 (3,850)	0.63 (4,508)
200 (100)	0.85 (1,881)	0.83 (6,921)	0.66 (8,786)
300 (150)	0.86 (2,457)	0.83 (9,938)	0.66 (13,211)
400 (200)	0.86 (2,873)	0.83 (12,526)	0.71 (17,259)
500 (250)	0.86 (3,247)	0.83 (15,065)	0.72 (21,460)
600 (300)	0.86 (3,633)	0.83 (17,479)	0.72 (25,558)

Table 6: AUC scores for “waiting time unacceptable” across training set sizes.

Figure 4: Lift charts for each service attribute.







In each chart, it is clear from the degree of “lift” observed, or the bump in each curve, that the specialized text analytic methods typically outperform both sentiment analysis (AFINN negative) and the random chance baseline. In each case, the unigram method provides the greatest AUC (see Table 5), although the bigram and trigram methods also frequently outperformed the baseline methods as well. While most of these charts illustrate similar trends, we did note a few unusual results, such as the poor performances of the “incident type emergency” bigrams and “environment unsafe/dirty” trigrams. A possible reason for the poor performance is that, due to the specificity of bigrams and trigrams relative to unigrams, overfitting occurs more frequently. Fortunately, in each case in which one method performed poorly, another method was observed to perform quite well.

Interestingly, although bigrams and trigrams typically performed worse than unigrams, they often offered similar levels of performance over a top-ranking set of reviews. However, due to the specificity of these terms, matching reviews were exhausted, and the remainder of reviews were classified at the rate of random chance. However, there is still some utility in these methods because the specificity of these terms may be a useful diagnostic tool, responding to the call by Greaves et al. (2013). For example, consider the unigrams generated versus the trigrams generated for “diagnosis misdiagnosed.” While unigrams such as “wrong” or “misdiagnosed” may reveal reviews concerning incorrect diagnoses quickly, they do not detail the nature of the misdiagnosis. Although trigrams may not yield as many matches, terms such as “a stroke and” or “the surgery after” may be more illuminating as to the nature of the diagnosis. Thus, although unigrams may provide superior performance, bigrams and trigrams may offer enhanced interpretability. The decision-maker may determine whether they prefer rapid sorting (unigrams) or more detailed insights (bigrams/trigrams).

6. Conclusions

Our results demonstrated a particular set of service attributes present in Facebook reviews of hospitals. Although these service attributes overlap somewhat with traditional metrics such as HCAHPS, they also offer a new perspective as to patients' views on service quality. These service attributes span the reviews very well, as 77.92% of the reviews mentioned at least one of the service attributes. Importantly, we showed that not all service attribute values affected overall ratings equally. In particular, communication, waiting time, and the helpfulness of treatments seemed to be highly associated with patients' star ratings. We also generated prevalent key terms for service attribute values, which can assist health care policy makers and health care providers to rapidly trace specific service concerns and modify their policies or resources to enhance their hospital services. This study highlights that these terms derived from service attributes are actionable and informative compared to the non-specific emotion-focused terms supplied by traditional sentiment analysis.

7. Limitations and Future Work

Our study is subject to several limitations. First, undergraduate students tagged reviews through manually in this study. When considering a complex medical case, the students may overlook reported service issues; for instance, some students may misinterpret reviews that actually discuss a medical error such as misdiagnosis by unknowingly claiming these reviews as only symptoms rather than tagging them as service concerns. However, the high degree of agreement with the authority tagger suggests that the tags are of generally high quality. Relatedly, the process of generating smoke terms is expensive in that it requires substantial labor as well as computational efforts. Although we did not pursue this extension, future work might examine the use of artificial intelligence to build a broad lexicon from a seed is also worth consideration in future work (Asghar, Ahmad, Qasim, Zahra,

& Kundi, 2016). Another possible limitation is the choice of sampling reviews from the top 100-listed hospitals in the United States. Hospitals outside the top 100-listed may share different performance characteristics. We cannot guarantee that our findings would apply to all hospitals. However, the top 100 hospitals, which this study focuses on, likely share many similar service and performance characteristics with a great variety of other hospitals. A final possible limitation is that we employed a tagging protocol at the review-level. That is, we asked the student taggers to indicate whether an entire review referred to a service attribute as opposed to the sentence-level or phrase-level. We expect that a more granular approach might reduce false positives, but it would also require an enormous tagging effort to address the same number of reviews with such specificity. Therefore, analysis at the review-level is quite common in the literature (Abrahams et al., 2015; Goldberg & Abrahams, 2018; Mummalaneni et al., 2018).

Gathering data from patients' online reviews implies that patients have self-selected their participation in online discussions. As such, it is crucial for healthcare administrators and policy makers to be aware that the findings are less likely to exemplify the whole population of the hospital service concerns. Instead, this discovery of specific service attributes should be viewed as supplementary to the conventional data collecting exercises (e.g. HCAHPS, patient surveys, etc.). The rapid discovery and investigation of the specific service attributes in this study is unique, and it would otherwise have been difficult for practitioners to analyze voluminous real-time (unstructured) data.

Future work may also look to categorize service-specific categories into leading, in-delivery, and trailing. In our study, we focused mostly on leading and in-delivery service attributes. A leading service attribute is one that occurs before the service occurs, while an in-delivery service attribute occurs while the service is occurring (e.g., doctor is talking with a patient). A trailing service attribute might be the medical outcome (e.g., successful treatment versus unsuccessful treatment).

The following questions could be addressed in future studies: would a patient that provides an initial low online rating feel the same way if one month later the treatment were to be deemed a success? In contrast, if a patient initially provided a high online rating but later experienced a serious side effect, would that be conducive to a poor rating? Gathering a rating before and after the service could help to assess these questions.

Future research may extend this text classification technique to web-based physician reviews (e.g., RateMDs, HealthGrades.com, etc.) to supply health care providers and researchers with substantial information that indicate patients' feedback of various medical fields. Additionally, these techniques may also have value in considering online consumer reviews of other products/services (Abrahams et al., 2015; Goldberg & Abrahams, 2018). Additionally, research studies could dig further in exploring the relationship between web-based ratings of hospitals and traditional types of hospital services acquired from the conventional survey results (e.g., HCAHPS).

Equipped with a stronger understanding of specific service attributes that influence online star ratings, managers will be better able to focus on key words and phrases of the service attributes that can improve the patient experience and the overall reputation of the hospital. If the online rating is strongly associated with a specific service attribute, then this identifies where administrators should focus to enhance the patient experience as opposed to other aspects with a lesser influence. In our study, communication, waiting times, and perceptions of treatment helpfulness seemed to be especially important factors. Before making policy changes, administrators should not only target the influential factor, but they should also focus on the cost of targeting each factor to make a final decision as to the most effective policy to implement.

References

- Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6), 975-990.
- Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, 54(1), 87-97.
- Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics*, 100, 108-120.
- Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016). SentiHealth: Creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, 5(1), 1139.
- Bardach, N. S., Asteria-Peñaloza, R., Boscardin, W. J., & Dudley, R. A. (2012). The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Quality & Safety*, bmjqs-2012-001360.
- Bardach, N. S., Lyndon, A., Asteria-Peñaloza, R., Goldman, L. E., Lin, G. A., & Dudley, R. A. (2015). From the closest observers of patient care: a thematic analysis of online narrative reviews of hospitals. *BMJ Quality & Safety*, bmjqs-2015-004515.
- BrightLocal (2016). Local Consumer Review Survey 2016.
- Brody, S., & Elhadad, N. (2010). Detecting salient aspects in online reviews of health providers. *AMIA Annual Symposium Proceedings: American Medical Informatics Association*. 202.
- Campbell, L., & Li, Y. (2017). Are Facebook user ratings associated with hospital cost, quality and patient satisfaction? A cross-sectional analysis of hospitals in New York State. *BMJ Quality & Safety*, bmjqs-2016-006291.
- Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345-354.
- Davis, Z. E. (2018). *Toward a Healthcare Services Ecosystem*, Virginia Tech.
- Devarakonda, M. V., Mehta, N., Tsou, C.-H., Liang, J. J., Nowacki, A. S., & Jelovsek, J. E. (2017). Automated problem list generation and physicians perspective from a pilot study. *International Journal of Medical Informatics*, 105, 121-129.
- Dobrzykowski, D. D., Callaway, S. K., & Vonderembse, M. A. (2015). Examining pathways from innovation orientation to patient satisfaction: A relational view of healthcare delivery. *Decision Sciences*, 46(5), 863-899.
- Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W., & Conway, M. (2016). Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annual Symposium Proceedings: American Medical Informatics Association*. 524.
- Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, 57(6), 74-81.

- Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems*, 40(2), 213-233.
- Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, 90, 1-11.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.
- Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, 2(Mar), 721-747.
- Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2014). Empowering difficult classes with a similarity-based aggregation in multi-class classification problems. *Information Sciences*, 264, 135-157.
- Glover, M., Khalilzadeh, O., Choy, G., Prabhakar, A. M., Pandharipande, P. V., & Gazelle, G. S. (2015). Hospital evaluations by social media: A comparative analysis of Facebook ratings among performance outliers. *Journal of General Internal Medicine*, 30(10), 1440-1446.
- Goldberg, D. M., & Abrahams, A. S. (2018). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*, 105, 52-65.
- Greaves, F., Lavery, A. A., Cano, D. R., Moilanen, K., Pulman, S., Darzi, A., & Millett, C. (2014). Tweets about hospital quality: a mixed methods study. *BMJ Quality & Safety*, 23(10), 838-846.
- Greaves, F., Pape, U. J., King, D., Darzi, A., Majeed, A., Wachter, R. M., & Millett, C. (2012). Associations between Web-based patient ratings and objective measures of hospital quality. *Archives of Internal Medicine*, 172(5), 435-436.
- Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, 15(11).
- Griffiths, A., & Leaver, M. P. (2018). Wisdom of patients: Predicting the quality of care using aggregated patient feedback. *BMJ Quality & Safety*, 27(2), 110-118.
- Hao, H., & Zhang, K. (2016). The voice of Chinese health consumers: A text mining approach to web-based physician reviews. *Journal of Medical Internet Research*, 18(5).
- Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., McIver, D. J., Rozenblum, R., Wright, A., & Bourgeois, F. T. (2015). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety*, bmjqs-2015-004309.
- Hu, N., Koh, N. S., & Reddy, S. K. (2014). 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, 42-53.
- Hu, N., Pavlou, P. A., & Zhang, J. J. (2017). On Self-Selection Biases in Online Product Reviews. *MIS Quarterly*, 41(2), 449-471.

- James, T. L., Calderon, E. D. V., & Cook, D. F. (2017). Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*, 71, 479-492.
- Johnson, D. M., & Russell, R. S. (2015). SEM of service quality to predict overall patient satisfaction in medical clinics: A case study. *Quality Management Journal*, 22(4), 18-36.
- Johnson, D. M., Russell, R. S., & White, S. W. (2016). Perceptions of care quality and the effect on patient satisfaction. *International Journal of Quality & Reliability Management*, 33(8), 1202-1229.
- Jung, Y., Hur, C., Jung, D., & Kim, M. (2015). Identifying key hospital service quality factors in online health communities. *Journal of Medical Internet Research*, 17(4).
- King, M. A. (2015). *Ensemble Learning Techniques for Structured and Unstructured Data*, Virginia Tech.
- Lagu, T., Goff, S. L., Craft, B., Calcasola, S., Benjamin, E. M., Priya, A., & Lindenauer, P. K. (2016). Can social media be used as a hospital quality improvement tool? *Journal of Hospital Medicine*, 11(1), 52-55.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.
- López, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: A qualitative content analysis. *Journal of General Internal Medicine*, 27(6), 685-692.
- Mummalaneni, V., Gruss, R., Goldberg, D. M., Ehsani, J. P., & Abrahams, A. S. (2018). Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety Science*, 104, 260-268.
- Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.
- Nottingham, Q. J., Johnson, D. M., & Russell, R. S. (2018). The effect of waiting time on patient perceptions of care quality. *Quality Management Journal*, 25(1), 32-45.
- Pai, R. R., & Alathur, S. (2018). Assessing mobile health applications with twitter analytics. *International Journal of Medical Informatics*, 113, 72-84.
- Paul, M. J., Wallace, B. C., & Dredze, M. (2013). What affects patient (dis) satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model. *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI*.
- Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., & Merchant, R. M. (2016). What can Yelp teach us about measuring hospital quality? *Health Affairs (Project Hope)*, 35(4), 697.
- Rastegar-Mojarad, M., Ye, Z., Wall, D., Murali, N., & Lin, S. (2015). Collecting and analyzing patient experiences of health care from social media. *JMIR Research Protocols*, 4(3).

Rodrigues, R. G., das Does, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, 85(1), 80-95.

Rosen, P. (2016). The patient as consumer and the measurement of bedside manner. *NEJM Catalyst*.

Rozenblum, R., & Bates, D. W. (2013). Patient-centred healthcare, social media and the internet: the perfect storm? BMJ Publishing Group Ltd.

Rozenblum, R., Greaves, F., & Bates, D. W. (2017). The role of social media around patient experience and engagement. *BMJ Quality & Safety*, 26(10), 845-848.

Thirumalai, S., & Sinha, K. K. (2009). Customization strategies in electronic retailing: Implications of customer purchase behavior. *Decision Sciences*, 40(1), 5-36.

Wagland, R., Richardson, A., Armes, J., Hankins, M., Lennan, E., & Griffiths, P. (2015). Treatment-related problems experienced by cancer patients undergoing chemotherapy: a scoping review. *European Journal of Cancer Care*, 24(5), 605-617.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6), 1098-1103.

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

Wolf, J. R., & Muhanna, W. A. (2011). Feedback mechanisms, judgment bias, and trust formation in online auctions. *Decision Sciences*, 42(1), 43-68.

Yang, M., Kiang, M., & Shang, W. (2015). Filtering big data from social media—Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, 54, 230-240.

Yang, P.-C., Lee, W.-C., Liu, H.-Y., Shih, M.-J., Chen, T.-J., Chou, L.-F., & Hwang, S.-J. (2018). Use of Facebook by hospitals in Taiwan: A nationwide survey. *International Journal of Environmental Research and Public Health*, 15(6), 1188.

Zhang, W., Deng, Z., Hong, Z., Evans, R., Ma, J., & Zhang, H. (2018). Unhappy Patients Are Not Alike: Content Analysis of the Negative Comments from China's Good Doctor Website. *Journal of Medical Internet Research*, 20(1).

Appendix A: Tagging protocol

For waiting time:

- Use “not mentioned” if the patient did not mention whether they were satisfied or not satisfied with the waiting time from the healthcare providers.
- Use “unacceptable” if the patient was dissatisfied with the waiting time for the healthcare provider (e.g. doctors, nurses, etc.).
- Use “acceptable” if the patient was happy with the waiting time from the healthcare providers.

For incident type:

- Use “not mentioned” if patient does not mention what their injury or condition was that resulted in their hospital visit.
- Use “emergency” if patient had a life-threatening condition (e.g. heart-attack or stroke). If patient subsequently received surgery, choose both “emergency” and “surgery.”
- Use “birth” if patient visited the hospital for childbirth (e.g. natural birth or caesarean section procedure). If Caesarean section, then choose both “birth” and “surgery”.
- Use “urgent” if patient had an injury or condition that was not life threatening (e.g. broken arm, sprained wrist, etc.). If patient subsequently received surgery, choose both “urgent” and “surgery”.
- Use “Surgery” if patient visited hospital for a scheduled surgery.

For diagnosis:

- Use “not mentioned” if patient did not mention whether their diagnosis by the hospital was correct or incorrect.
- Use “correctly diagnosed” if patient mentioned that the hospital diagnosed them correctly.

- Use “misdiagnosed” if patient mentioned that the hospital misdiagnosed their illness (i.e. doctor / nurse / hospital gave the wrong diagnosis).
- Use “undetermined” if patient mentioned that the hospital / doctor / nurse could not find the reason for their illness.

For treatment:

- Use “not mentioned” if patient did not mention the outcome of the treatment or intervention provided by the hospital.
- Use “helpful” if patient mentioned that their health improved as a result of the hospital treatment / intervention.
- Use “harmful” if patient mentioned that their health was worse as a result of the hospital treatment / intervention. Include medication errors, poor procedures (e.g. failure to properly clean wounds, resulting in infection, etc.).

For communication:

- Use “not mentioned” if the patient did not mention their interaction experience with the doctors, nurses, or staff.
- Use “bad” if patient indicates that there is a lack of interaction between the patient and healthcare providers (e.g. doctors, physicians, nurses, staff members), or patients experienced rude behavior from hospital workers and healthcare providers.
- Use “good” if patient indicates that healthcare providers (e.g. doctors, physicians, nurses, staff members) have active interaction with the patient, or patients experienced friendly/polite behavior from hospital workers and healthcare providers

For environment:

- Use “not mentioned” if patient does not mention the sanitation/hygiene/cleanliness of the hospital rooms or other materials.
- Use “clean / sanitary” if the patient comments hospital environment as being clean hospital rooms and other materials, e.g. clean beddings, clothes, sanitized hospital equipment, etc.
- Use “dirty / unsanitary” if the patient comments hospital environment as being dirty hospital rooms and other materials, e.g. dirty beddings, clothes, not sanitized hospital equipment, etc.

For cost:

- Use “not mentioned/satisfactory” if patient did not mention the cost of the procedure / intervention or was satisfied with the cost.
- Use “too expensive” if patient was unhappy with the costs / charges levied by the hospital.

Appendix B: Inter-rater reliability metrics

Table 7 shows inter-rater reliability metrics for each service attribute assessed. Comparing the 300 tags of the lead researcher with the most conservative tag by student taggers on matching reviews yields the reported percentage agreement. Cohen's κ compares this agreement percentage to random chance, where values closer to 1 indicate strong agreement relative to random chance. All Cohen's κ values were greater than 0.4, indicating “moderate” or better agreement” (Landis & Koch, 1977) or “fair to good” or better agreement (Fleiss, 1971) between coders.

Service attribute	Percentage agreement	Cohen's κ
Waiting time	92.89%	0.89
Incident type	75.89%	0.70
Diagnosis	59.68%	0.46
Treatment	69.17%	0.53
Communication	71.54%	0.57
Environment	95.26%	0.93
Cost	97.23%	0.95

Table 7: Inter-rater reliability metrics for each service attribute.

Next, we can compute Cohen's κ for each service attribute value by treating that service attribute as a binary variable. For example, consider the “waiting time” service attribute, which contains three possible values. To compute Cohen's κ for the “acceptable” value, we treat all “acceptable” tags as 1 and all other tags as 0; then, we can assess inter-rater agreement for this specific service attribute value. This analysis is displayed in Table 8. All Cohen's κ values except for diagnosis “not mentioned” were greater than 0.4, again indicating “moderate” or better agreement” (Landis & Koch, 1977) or “fair to good” or better agreement (Fleiss, 1971) between coders.

Service attribute	Service attribute value	Percentage agreement	Cohen's κ
Waiting time	Acceptable	98.02%	0.96
Waiting time	Unacceptable	94.07%	0.88
Waiting time	Not mentioned	93.68%	0.87
Incident type	Urgent	93.68%	0.87
Incident type	Emergency	93.68%	0.87
Incident type	Surgery	91.70%	0.83
Incident type	Birth	95.26%	0.91
Incident type	Not mentioned	72.73%	0.45
Diagnosis	Correct	70.36%	0.41
Diagnosis	Misdiagnosed	94.86%	0.90
Diagnosis	Undetermined	92.09%	0.84
Diagnosis	Not mentioned	62.06%	0.24
Treatment	Helpful	73.52%	0.47
Treatment	Harmful	94.07%	0.88
Treatment	Not mentioned	70.75%	0.42
Communication	Good	77.47%	0.55
Communication	Bad	93.28%	0.87
Communication	Not mentioned	72.33%	0.45
Environment	Clean/sanitary	96.84%	0.94
Environment	Dirty/unsanitary	98.42%	0.97
Environment	Not mentioned	95.26%	0.91
Cost	Too expensive	97.23%	0.94
Cost	Not mentioned/satisfactory	97.23%	0.94

Table 8: Inter-rater reliability metrics for each service attribute value.

Appendix C: Most prevalent terms for each service attribute value.

Table 9 shows the 30 most prevalent 1-word, 2-word, and 3-word terms (phrases) for each service attribute value are shown below; **bold words** reflect words that overlap with words from the past studies; *italicized words* and underlined words indicate negative sentiment words and positive sentiment words respectively from the AFINN dictionary; remaining words indicate non-sentiment words found in our study that are distinct from past studies.

Service attribute value	Type of term	Top terms curated
Waiting time: unacceptable	Words	hours , waiting , wait , room, er, pm, sitting, hour , <i>ridiculous</i> , waited, doctor , slow, <i>worst</i> , <i>horrible</i> , finally, sit, left , emergency, <i>pain</i> , blood , sat, hallway, appointment , <i>unacceptable</i> , minutes, call, chest, bed , busy, checked
Waiting time: unacceptable	2-word phrases	waiting room, the waiting , the er, waiting for, a room, still waiting , hours and, 6 hours , 5 hours , to wait , waiting in, hours in , hours for, been waiting , sitting in, room for, be seen, 3 hours , no one, waiting to, an hour , the doctor, wait for, 4 hours, is <i>ridiculous</i> , a bed, and still, er for, hours to, two hours
Waiting time: unacceptable	3-word phrases	in the waiting , the waiting room, in the er, to be seen, waiting in the, sitting in the, and still waiting , this is <i>ridiculous</i> , for a room, hours in the, for 4 hours , waiting room for, the <i>worst</i> experience, the er for, still waiting to, for a bed, for 5 hours , have been for, been waiting in, er for over, see the doctor, hours of waiting , to wait for, emergency room for, waiting for a, in the hallway, been here since, have to wait , the emergency room, hours with my
Communication : bad	Words	rude , <i>worst</i> , told, room, <i>horrible</i> , call , <i>pain</i> , doctor , unprofessional , left , nurse , back, blood , patients , trying, come, called, phone, money , <i>disappointed</i> , <i>poor</i> , anything, <i>terrible</i> , appointment , said, leave, patient , see, hour , sitting

Communication : bad	2-word phrases	no one, the worst , the nurse , the doctor , rude and, to get, worst hospital, a nurse , trying to, a doctor , told me, worst experience, the room, a room, get a, <i>disappointed</i> with, was told, are rude , customer service, to call, lack of, come to, i told, very rude , supposed to, pain and, the patient , told that, tell the, pain meds
Communication : bad	3-word phrases	the worst hospital, the worst experience, are rude and, i was told, to this hospital, to speak to, we were told, worst hospital ever, was supposed to, in the room, is the worst , very <i>disappointed</i> with, to use the, never go back, very disappointed with, to use the, never go back, <i>disappointed</i> with the, the lack of, into a room, to get back, do their job, i told her, hours to be, to see the, was told that, worst experience ever, told me that, trying to get, cant do anything
Treatment: harmful	Words	pain , blood , meds, worst , lies, denied , <i>die</i> , ambulance, vomiting, genital, home, went, legs, <i>hurting</i> , ethics, unsanitary, left , medicine, never, get, infection, mutilation, medication, unable, urine, bed, condition, genitals, cut, bite
Treatment: harmful	2-word phrases	pain medication, the pain , did not, a blood , genital mutilation, they let, much pain , they sent, gave me, the worst , her home, days later, 1 star, home with, went to, to get, no harm, i decided, an ambulance, of chest, was trying, her meds, the <i>horrible</i> , placed on, my throat, a spider, they send, sent him, to <i>kill</i> , never did
Treatment: harmful	3-word phrases	and they sent, i decided to, do no harm, on the floor , sent her home, work for the, have an emergency, the worst hospital, staff was rude , so much pain , to stop the, in the ambulance, the middle of, in the middle, was very <i>disappointed</i> , nurse came in, a blood clot, i was left, bedside manners and, its a shame, wait for a, the genitals of, in an emergency, blood all over, saline solution but, very upset with, hospital less than, not care about, who would <u>love</u> , hospital treated my
Environment: dirty	Words	dirty , filthy , bathrooms, room, tables, floors , cleaned, wet, highlights, stinks, restrooms, seats, renovations, sheets, brushed, nastiest, overflowing, banging, hot, urine, garbage, brown, pictures, food, bed , <i>dissatisfied</i> , ground, ratchet, unsanitary, quiet
Environment: dirty	2-word phrases	<i>dirty</i> and, is filthy, dirty it, was filthy, is <i>dirty</i> , the floor , rude staff, room was, was gross, its filthy, hot and, clean it, dont look, rude nurses, was <i>dirty</i> , the nastiest, was pleasant, filthy and, dirty rooms, bathrooms are, <i>dirty</i> room, <i>horrible</i> staff, blood on, the ground, place was, showed up, looked at, hot water, could care, check my

Environment: dirty	3-word phrases	hospital is dirty, dont look like, made a huge, on the ground, room was filthy, emergency room was, they could <u>care</u> , blood on the, no hot water, could <u>care</u> less, <i>worst</i> place ever, had the <i>worst</i> , faces are showing, throat felt like, admitted and still, my families hospital, the floor had, i could bear, patient advocacy is, did housekeeping and, brushed off this, schedule their patients, tested <u>amazing</u> surgeries, avoid at all, <i>problem</i> nurses are, <i>dissatisfied</i> the overnight, <i>dirty</i> blood stain, randomly move to, just stopping by, exaggerating 2 bathrooms
Diagnosis: misdiagnosed	Words	<i>pain</i> , wrong , misdiagnosed , tylenol, muscle, meds, conscious, refusing, numb, explanation, <i>faking</i> , evaluated, misdiagnosing, tests , turns, infected, dropped, cause, blood , <i>kill</i> , move, diagnosis, amount, doctor , ambulance, release, severity, gurney, chills, releasing
Diagnosis: misdiagnosed	2-word phrases	<u>fine</u> and, <i>wrong</i> with, home with, was <i>wrong</i> , was <u>fine</u> , the <i>pain</i> , reconsider your, they misdiagnosed, <i>losing</i> my, surgery after, normal healthy, and <i>refusing</i> , wouldnt treat, a conscious, <i>pain</i> you, his wrist, <i>kill</i> my, of fluid, dropped the, discharged with, conscious man, <i>pain</i> meds, blood and, an ambulance, i ended, chest <i>pain</i> , found out, sent to, turns out, please reconsider
Diagnosis: misdiagnosed	3-word phrases	sent me home, home with a, please reconsider your, a stroke and, week later we, tried to <i>kill</i> , night due to, multiple times and, <i>pain</i> and the, me home with, he was <u>fine</u> , to <i>kill</i> my, the surgery after, was <u>fine</u> and, something was <i>wrong</i> , a conscious man, <i>pain</i> you are, call her back, in his heart, and <i>refusing</i> to, not go here, they dont tell, <i>kill</i> my dad, to call her, doctors and they, reconsider your decision, with a bunch, what was happening, on <i>pain</i> meds, help me at
Cost: too expensive	Words	pay, bill, money , billing, insurance , billed , card, charged , afford, paying , cost , credit, payment, paid, pocket, <i>illegal</i> , expensive, <i>scam</i> , balance, fees, account, collection, deductible, <i>demand</i> ed, tickets, overcharge, sum, collections, thousands, office
Cost: too expensive	2-word phrases	to pay, pay the, insurance card, afford to, billing dept, of pocket, a bill, the billing, co pay, billing department, pay for, not cover, my credit, you pay, <i>not</i> afford, for services, my account, paid out, paying for, and pay, the money, to bill, insurance they, <i>not</i> paying, and money, bill for, facility fees, <i>demand</i> ed my, too expensive, a copay

Cost: too expensive	3-word phrases	out of pocket, 0 co pay, they continue to, can <i>not</i> afford, to pay for, the billing payment, time and money, had to pay, room visit and , considered a emergency, money they are, <u>care</u> is actually, insurance did not, the billing dept, wanted to self, the agreed upon, my credit score, when i returned, insurance they <i>demanded</i> , self pay even, my insurance card, a emergency room, have insurance they, trying to collect, card since i, a receipt from, im not paying, not afford to , was insured and, great but they
Incident type: urgent	Words	fever, cancer, pain , doctor , sponge, arm , bleed, liquid, infection, shifts, injection, narcotics, prescribed, asthma, antibiotics, pills, dehydrated, stomach, came, clue, head, ended, hours , filled, water, seizures, addict, slipped, placement, abscess
Incident type: urgent	2-word phrases	pain and, the doctor, diagnosed with, a fever, see her, a drug, saturday morning, filled out, my chart, my arm, so <i>bad</i> , and started, left me, ended up, i guess, and took, fever and, sponge bath, no clue, on drugs, been caring, and <i>lazy</i> , in care, another doctor, of <i>pain</i> , doctors have, a sponge, to weak, drug addict, the lung
Incident type: urgent	3-word phrases	i took my, get ahold of, came to check, 1 hour and, has been there, could go home, hospital to take, for <u>helping</u> me, was diagnosed with, to the nursing, seen by the, the doctor came, been seen by, and found out, came in and, room and she, on drugs she, goes wrong ill, checked in on, filth for that, her doctors have, hospital so doped, dehydrated and something, the patient but, in a million, nobody came to, i cant stand, got there at, and they wouldnt, up on drugs
Incident type: emergency	Words	er, <i>emergency</i> , room , pain , chest, went, <i>accident</i> , sent, ambulance, rushed, doctor , urine, seen, took, <i>attack</i> , <i>horrible</i> , nurse, admitted, lifted, <i>furios</i> , came, now, burn, arrived, life, brought, pressure, ridiculous , laying, <i>died</i>
Incident type: emergency	2-word phrases	the er, emergency room, the emergency, in er, an emergency, a room, i went, this er, er was, went to, er is, er room, er doctor, er staff, by ambulance, heart attack, emergency surgery, er on, in emergency, was rushed, er for, air lifted, <i>worst</i> emergency, <i>worst</i> er, look at, admitted into, too busy, for saving, your er, to er

Incident type: emergency	3-word phrases	in the er, the emergency room, to the er, the waiting room, waiting in the, in the emergency, at the er, <i>worst</i> emergency room, the er on, the er at, emergency room and, the er is, for an emergency, in severe pain, an emergency room, the er staff, was admitted into, admitted into the, have a room, into the er, to this er, the er now, a kidney store, was air lifted, emergency room experience, his blood pressure, laying on a, the <i>worst</i> emergency, into the emergency, on a stretcher
Incident type: surgery	Words	surgery, transplant, surgeon, heart, brain, recovery, open, surgical, surgeries, liver, kidney, <u>thanks</u> , team, surgeons, <u>thank</u> , performed, life, <u>amazing</u> , <u>positive</u> , pancreas, spine, hip, back, repair, post, complications, replacement, months, scheduled, nurses
Incident type: surgery	2-word phrases	surgery and, heart surgery, open heart, surgery here, my surgery, had surgery, back surgery, for surgery, liver transplant, kidney transplant, brain transplant, a liver, surgery there, a kidney, the surgery, surgery was, surgery in, surgery at, had open, got my, after surgery, my open, my life, a surgery, hip replacement, heart transplant, my surgeon, surgery to, surgery the, surgery is
Incident type: surgery	3-word phrases	open heart surgery, a liver transplant, had open heart, a kidney transplant, i had surgery, had a kidney, my open heart, had surgery there, had a liver, had surgery there, had heart surgery, had back surgery, to the team, and doctors that, replacement surgery and, surgery and the, heart surgery here, surgery here and, for the wonderful, asked for better, i got my, to have surgery, i had major, doctors and surgeons, after the surgery, before and after, wanted to stay, the clinic in, for surgery and, staff there was
Incident type: birth	Words	born, baby , birth, delivery , labor, delivered , babies, section , pregnancy, maternity, midwives, pregnant , twins, birthing, breastfeed, twin , experience , deliver, premature, delivering, weeks , <i>risk</i> , prenatal, kids, child, ward, partum, formula, csection, grandchild
Incident type: birth	2-word phrases	was born, born here, my baby , gave birth, birth to, labor and, and delivery , c section, i delivered, delivered my, born at, a baby , i gave, born there, were born, in labor, baby here, my pregnancy, our baby , the maternity, delivery and, my delivery , my babies, emergency c, maternity ward, first child, my twin, the baby , to deliver, baby was

Incident type: birth	3-word phrases	was born here, gave birth to, had my baby , birth to my, i gave birth, i was born, labor and delivery , i delivered my, was born at, both of my, i had my, were born here, was born there, via c section, have a baby , and my baby , emergency c section, had both of, c section on, a baby in, born here and, delivered my baby , my baby the, the maternity ward, the birth of, moment i arrived, with the midwives, a million times, baby here and, my first child
---------------------------------	---------------------------	---

Table 9: Most prevalent terms for each service attribute value.

Conclusion, Implications, and Future Work

The conclusion and several implications for stakeholders of each chapter are summarized below along with suggestions for future studies.

In *Chapter 1*, we used different text-classification methods to segment the safety concern reports from an enormous volume of consumer-generated reviews across a diverse set of product categories. We observed that a small portion of our tagged dataset across multiple product categories actually referred to safety concerns. Our text analytics approach suggests terms that are indicative of safety concerns in these online reviews within the particular event, hazard, and injury categories. We used unsupervised topic modeling techniques to reveal six latent topics that were related only to hazard categories present in these reviews. However, by using a supervised learning technique, we provided terms that practitioners and researchers could use to rapidly categorize the safety concern reports based upon event, hazard, and injury categories. Doing so adds specificity to safety surveillance and allows both firms and regulators to analyze online reviews for safety concern reports in greater detail. Additionally, we discovered that the sentiment approach was less effective than the supervised learning mechanism in predicting category-specific safety concerns. Hence, category-specific safety concern terms were able to isolate safety concerns quite reliably. Finally, after assessing the supervised classification methods on one dataset, we also used another dataset taken from a different source (i.e., the Fortune 50 retailer) for evaluating the generalizability of these classification methods.

Our findings have important implications for safety surveillance. In terms of implication for the practitioners, we found that online reviews are a substantial source of safety concern-related feedback across multiple product categories. However, we also found substantial variability across product categories in the proportion of reviews that referred to category-

specific safety concerns. Due to this reason, sentiment analysis cannot be relied upon to monitor particular safety concerns effectively in most product categories. While sentiment analysis continues to be a valuable tool for detecting the customer's emotive opinions about products, it is not the most efficient way for the practitioners to identify category-specific safety concerns across multiple product categories. On the other hand, our results prove that the smoke-scoring methodology substantially discovers more non-emotive opinions about products than sentiment analysis. These non-emotive narratives are prevalent in mentioning category-specific safety concerns. This study highlights that the safety concern-specific terms derived from safety concern categories retrieve more actionable insights of product safety than the non-specific emotion-focused terms supplied by traditional sentiment analysis. Therefore, by using our safety concern-specific smoke term dictionaries (generated in this study), the practitioners should take advantage of rapidly identifying, categorizing as well as prioritizing the specific safety concern reports from online reviews.

Next, we demonstrated in our study that, although different product categories have distinct safety concerns, there are some commonalities in hazard categories across product categories. These commonalities may be difficult to observe in the smaller dataset, but using a large dataset of over 120,000 reviews, we were able to collect enough verified safety concern reports to reveal a wide range of hazard categories by using the supervised classification method. In terms of research implications, we should gather substantially larger datasets for low-incident target classes (Event: injury mitigation, treatment; Hazard: biological; Injury: electric shock, foreign bodies, fracture, head injury, poisoning), in order to train more accurate smoke term dictionaries for these rare target classes. While these target classes were rare in our dataset, they are nevertheless critically important to safety hazard discovery, and thus deserve further, focused

data gathering effort, to acquire additional consumer-generated narratives describing these target classes. Consumer-generated narratives are especially important as prior studies (Winkler et al., 2016) indicate that narratives generated by medical professionals (e.g. as provided in the CPSC NEISS dataset) have limited efficacy in detecting safety concerns expressed by consumers in online reviews, due to the different lingo used by these stakeholder groups.

Hence, after building safety concern-specific smoke term lists, we will develop more robust assessments of the safety concern-specific smoke term lists, especially for the low-incident target classes. Therefore, our results should provide sets of unique terms that are useful in identifying, categorizing and prioritizing category-specific safety concerns reports across this broad spectrum of product categories. Relatedly, we demonstrated that safety surveillance performed on a more granular scale than considered in prior work. We showed this facet through our observed safety concern categories, and we further aim to enhance our risk assessment framework for analyzing the risk of varied products in our future studies. We will pursue future research by using supervised machine learning to filter these safety concern reports from online reviews on the basis of the event, hazard, and injury categories for the low-incident target classes, allowing organizations to determine risk scores of products on the basis of this further level of categorization. Ideally, we will construct a model that aggregates multiple categories of safety concern reports and uses risk assessment criteria to prioritize the most serious safety concerns.

In **Chapter 2**, we developed a model for determining the risk level of each product and exploring the text-classification of hazard code types. We tested our risk assessment model after conducting the semi-automatic categorization of sentences using online reviews of products across multiple product categories. In addition, we included risk factors by adapting them from

practitioners' risk-assessment frameworks and integrating online reviews with the text analytics to determine the computerized risk scores (i.e., likelihood of a hazard, the severity level, and the overall risk level, of a product). These risk scores helped us map the risk level of a specific product on the risk assessment matrix supplied in the EU guideline³⁵. The semi-automated framework not only considered the content of the online (real-time) reviews but also utilized the gathering of raw counts of customer reviews from Amazon's online data source. Ultimately, the risk level matrix should help support decision making for stakeholders like compliance officers, risk-management officers, and other risk-prevention experts. Therefore, the framework will help them decide whether they should mitigate the risks or hazards related to a specific model of their product.

To enlarge our sample size of hazard reports, we plan to merge multiple data sources in addition to Amazon.com for our future study. For example, Walmart, Target, and eBay.com also contain a large pool of online reviews for consumer products. Executing the semi-automatic classification procedures on a sample size gathered from multiple sources will increase the chance of collecting a high number of unique hazard reports. This larger sample size will help build a more robust method to help businesses find, classify, and prioritize risks at the *category*, as well as the product level.

As this study focused on achieving a few goals listed by the CPSC strategic plan in *Chapter 2*, we plan to modify our existing methodologies on accomplishing one or more remaining goals in our future study. For example, one of these remaining goals includes implementing hazard surveillance of consumer products to discover recalled products. Future

³⁵ Union, E. Guidelines for the management of the European Union Rapid Information System 'RAPEX'. Retrieved from: https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/repository/content/pages/rapex/docs/Guidelines%20annex_en.pdf

work could develop the automatic text classification approach in detecting recalled products that are sold in electronic marketplaces. As a result, safety regulators could potentially more rapidly identify sellers who are selling the recalled products online to consumers.

In **Chapter 3**, our results demonstrated a particular set of service attributes present in Facebook reviews of hospitals. Although these service attributes overlap somewhat with traditional metrics such as HCAHPS, they also offer a new perspective of patients' views on service quality. These service attributes span the reviews very well, as 77.92% of the reviews mentioned at least one of the service attributes. Most importantly, we showed that not all service attribute values affected overall ratings equally. In particular, communication, waiting time, and the helpfulness of treatments seemed to be highly associated with patients' star ratings. We also generated prevalent key terms for service attribute values, which can assist healthcare policymakers and healthcare providers to rapidly monitor specific service concerns and modify their policies or resources improve hospital services. This study highlights that these terms derived from service attributes are actionable and informative compared to the non-specific emotion-focused terms supplied by traditional sentiment analysis.

Our study is subject to several limitations: First, undergraduate students tagged reviews manually in this study. When considering a complex medical case, the students may overlook reported service issues; for instance, some students may misinterpret reviews that actually discuss a medical error such as misdiagnosis by unknowingly claiming these reviews as only symptoms rather than tagging them as service concerns. However, the high degree of agreement with the authority tagger suggests that the tags are of generally high quality. Relatedly, the process of generating smoke terms is expensive in that it requires substantial labor as well as computational efforts. Although we did not pursue this extension, future work might examine the

use of artificial intelligence to build a broad lexicon from a seed (Asghar, Ahmad, Qasim, Zahra, & Kundi, 2016). Another possible limitation is the choice of sampling reviews from the top 100-listed hospitals in the United States. Hospitals outside the top 100-listed may share different performance characteristics. We cannot guarantee that our findings would apply to all hospitals. However, the top 100 hospitals, which this study focuses on, likely share many similar service and performance characteristics with a great variety of other hospitals. A final possible limitation is that we employed a tagging protocol at the review level. That is, we asked the student taggers to indicate whether an entire review referred to a service attribute as opposed to the sentence-level or phrase-level. We expect that a more granular approach might reduce false positives, but it would also require an enormous tagging effort to address the same number of reviews with such specificity. Analysis at the review-level is quite common in the literature (Abrahams et al., 2015; Goldberg and Abrahams, 2018; Mummalaneni et al., 2018).

Gathering data from patients' online reviews implies that patients have self-selected their participation in online discussions. As such, it is crucial for healthcare administrators and policymakers to be aware that the findings are less likely to exemplify the whole population of the hospital service concerns. Instead, this discovery of specific service attributes should be viewed as supplementary to the conventional data collecting exercises (e.g. HCAHPS, patient surveys, etc.). The rapid discovery and investigation of the specific service attributes in this study are unique, and it would otherwise have been difficult for practitioners to analyze voluminous real-time (unstructured) data.

Future work may also look to categorize service-specific categories into leading, in-delivery, and trailing. In our study, we focused mostly on leading and in-delivery service attributes. A leading service attribute is one that occurs before the service occurs, while an in-

delivery service attribute occurs while the service is occurring (e.g., a doctor is talking with a patient). A trailing service attribute might be the medical outcome (e.g., successful treatment versus unsuccessful treatment). The following questions could be addressed in future studies: would a patient that provides an initial low online rating feel the same way if one month later the treatment were to be deemed a success? In contrast, if a patient initially provided a high online rating but later experienced a serious side effect, would that be conducive to a poor rating? Gathering a rating before and after the service could help to assess these questions.

Future research may extend this text classification technique to web-based physician reviews (e.g., RateMDs, HealthGrades.com, etc.) to supply health care providers and researchers with substantial information that indicate patients' feedback of various medical fields. Additionally, these techniques may also have value in considering online consumer reviews of other products/services (Abrahams et al., 2015; Goldberg and Abrahams, 2018). Additionally, research studies could dig further into exploring the relationship between web-based ratings of hospitals and traditional types of hospital services acquired from the conventional survey results (e.g., HCAHPS).

Equipped with a stronger understanding of specific service attributes that influence online star ratings, managers will be better able to focus on key words and phrases of the service attributes that can improve the patient experience and the overall reputation of the hospital. If the online rating is strongly associated with a specific service attribute, then this identifies where administrators should focus to enhance the patient experience as opposed to other aspects with lesser influence. In our study, communication, waiting times, and perceptions of treatment helpfulness seemed to be especially important factors. Before making policy changes, administrators should not only target the influential factor, but they should also focus on the cost of targeting each factor to make a final decision as to the most effective policy to implement.

In summary, this dissertation has presented novel text analytics methods for detecting key product quality concerns (Chapter 1 and 2) and service quality concerns (Chapter 3) in online reviews, using online retailers (Chapter 1 and 2) and hospitals (Chapter 3) as case studies. Our methods have centered on sub-classifying product and service quality concerns into key concern subtypes, and building and assessing special-purpose smoke term dictionaries for each key concern subtype, to enable automated classification at the sentence or full review level. The ability to automatically detect key concern subtypes is then leveraged to estimate the influence of the key concern subcategories – e.g. influence on product risk (Chapter 2), or influence on patient satisfaction (Chapter 3). Our assessments indicate these new models are helpful, though limited. Future work will tackle model restructuring and enhancement.