Decision Support Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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

Toy safety surveillance from online reviews

Matt Winkler [a], Alan S. Abrahams [a,*], Richard Gruss [a], Johnathan P. Ehsani [b]

[a] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, Pamplin Hall, Suite 1007, 880 West Campus Drive, Blacksburg, VA 24061, United States
[b] Center for Injury Research & Policy, Johns Hopkins Bloomberg School of Public Health, Hampton House, Room 554, 624 N. Broadway, Baltimore, MD 21205, United States

ARTICLE INFO

Article history:
Received 4 December 2015
Received in revised form 20 June 2016
Accepted 22 June 2016
Available online xxxx

Keywords:
Online reviews
Safety surveillance
Toys
Injuries

ABSTRACT

Toy-related injuries account for a significant number of childhood injuries and the prevention of these injuries remains a goal for regulatory agencies and manufacturers. Text-mining is an increasingly prevalent method for uncovering the significance of words using big data. This research sets out to determine the effectiveness of text-mining in uncovering potentially dangerous children's toys. We develop a danger word list, also known as a "smoke word" list, from injury and recall text narratives. We then use the smoke word lists to score over one million Amazon reviews, with the top scores denoting potential safety concerns. We compare the smoke word list to conventional sentiment analysis techniques, in terms of both word overlap and effectiveness. We find that smoke word lists are highly distinct from conventional sentiment dictionaries and provide a statistically significant method for identifying safety concerns in children's toy reviews. Our findings indicate that text-mining is, in fact, an effective method for the surveillance of safety concerns in children's toys and could be a gateway to effective prevention of toy product-related injuries.

1. Introduction

In 2011, a child was treated in a U.S. emergency department for a toy-related injury every 3 min [2]. As a result, toy injuries are of major concern to various stakeholders, including toy manufacturers and parents of children who play with these toys. The NPD Group, a market research company that tracks about 80% of the U.S. toy retail market, determined that the toy market consisted of $18.11 billion of sales in 2014, a 4% increase from the 2013 number of $17.47 billion [33]. The toy categories with the highest annual sales were: action figure/accessories/role play, arts and crafts, building sets, dolls, games/puzzles, infant/preschool, youth electronics, outdoor and sports toys, plush, and vehicles.

The U.S. Consumer Product Safety Commission issued a total of 401 toy recalls in the seven fiscal years from 2008 to 2014, resulting in significant expenses to toy manufacturers, retailers, and consumers [41]. According to the U.S. Consumer Product Safety Commission's *Toy-Related Deaths and Injuries, Calendar Year 2013* report, there were an estimated 256,700 toy-related injuries treated in the U.S. in 2013 [40], 73% of these injuries occurred to children younger than 15 years of age, 69% to children younger than 12, and 33% to children younger than 5.

A recent example of a children's toy that was recalled was the "My Sweet Love/My Sweet Baby Cuddle Care Baby Doll". Walmart recalled 174,000 of these dolls due to a burn hazard [30]. The CPSC reported that a circuit board in the doll's chest could overheat, causing the surface of the doll to burn the user of the product [30]. Walmart received 12 incident reports which included two burns or blisters to the thumb. The CPSC advised consumers to stop using this product and immediately return the doll to any Walmart store for a refund. In separate toy recall cases, reported in the *New York Times*, Mattel recalled over eighteen million toys due to lead paint hazards, and due to the risk of small powerful magnets being swallowed [35].

The United States Consumer Product Safety Commission (CPSC) oversees the toy industry. In 2008, the Consumer Product Safety Improvement Act (CPSIA) provided the CPSC with new regulatory and enforcement powers to enhance several CPSC statutes [19]. The CPSIA maintained a particular focus on classification and regulation of children's products. The CPSC both tests toys and responds to reports of incidents in order to enforce safety violation standards. The CPSC has jurisdiction over 15,000 types of products, with toys consisting of a small portion of this jurisdiction. In 2015, the CPSC had about 500 employees directed at hazard identification and reduction. With 3000 to 5000 new toys being introduced by toy manufacturers each year, the CPSC is unable to police or test every toy and often responds to a safety issues after they have already occurred. As a result of these resource constraints, plenty of dangerous toy products arrive at stores every year. Many toy companies test their products in their own labs before offering the products to the public, but there remain a significant number of toys that are not tested. We believe that a vast amount of useful text data embedded in millions of online consumer reviews can be utilized by toy manufacturers, parents, and the CPSC, to advance safety surveillance.

* Corresponding author. Tel.: +1 540 231 5887; fax: +1 540 231 3752.
E-mail addresses: winkmat7@vt.edu (M. Winkler), abra@vt.edu (A.S. Abrahams), rgruss@vt.edu (R. Gruss), jpehsani@gmail.com (J.P. Ehsani).

http://dx.doi.org/10.1016/j.dss.2016.06.016
0167-9236/

Consumers rely heavily on the Internet for information about product safety and reliability, including children's toys. Consumers provide their manufacturers, sellers, and their fellow consumers with information about product safety and reliability through sources such as product reviews on retailer websites. Manually identifying and analyzing consumer reviews among millions of consumer postings that relate to product safety issues is a challenging task. Using text-mining to identify and prioritize the vast volume of online reviews regarding safety issues in children's toy products is the focus of this paper.

The rest of this paper is structured as follows. First, we motivate the need for quality surveillance research targeted specifically at the discovery of safety concerns from textual online discussion forums. Next, we discuss and contrast related work. We describe our contributions and the research questions we aim to address. We lay out a process for quality surveillance in the toy industry using analysis of online reviews, recalls, and injury reports. We evaluate our safety issue discovery approach using three experiments on a large sample data set. Finally, we draw conclusions and propose future work.

2. Background and related work

In this section, we review related work in sentiment analysis, online reviews, text mining, and social media surveillance, and explain their relationship to children's toy issue surveillance. We review the coverage and limitations of prior work, as well as the research questions raised. The research gaps associated with past studies are discussed to highlight how these methods could be improved upon.

2.1. Sentiment analysis

Sentiment analysis refers to natural language processing techniques used to quantify the type and amount of emotion expressed in text. Common dictionary sources for sentiment analysis, such as the AFINN [31], ANEW [9], and Harvard General Inquirer [23] dictionaries, assign scores or categories to words in order to assess sentiment. The SentiStrength [37,38] and OpinionFinder [42,43] sentiment analysis methods go beyond basic sentiment scoring techniques—which use constant word scores irrespective of word context—and provide for more complex, context-aware sentiment determination.

In online product reviews, a sentence or review with net positive sentiment score is taken to indicate praise of a particular product and a net negative score indicates criticism of a product. Abbasi, Chen, and Salem analyzed linguistic data in online discussion forums to quantify opinions of users [1]. Other studies have applied sentiment analysis to predict a firm's earnings and returns [26,36], the directional movement of a firm's stock price [32], or its market volatility [6].

Sentiment analysis can be a useful tool in uncovering consumer opinions regarding products, including the children's toy industry. Accessing sources such as online reviews to uncover consumer concerns, using negative scores, and consumer satisfaction, using positive scores, can provide toy manufacturers as well as regulatory agencies with useful information. However, there are limitations involved in using sentiment analysis. Firstly, the most basic sentiment analysis techniques, which use single-word markers, are not always effective in determining positive and negative tones. For example, a consumer could provide a review of a children's toy, stating "This toy is not *bad* at all, my two-year old plays with it all the time." The word *bad*, viewed alone, is assigned a negative sentiment score, leading the researcher to believe that this review was negative when in fact it revealed a positive opinion of this particular toy product. Secondly, many sentiment analysis approaches are generic and domain-independent, so domain-specific danger-words may not be recognized: consider the word "recall" which, in its typical connotation of "remember" (e.g. "I recall the time…"), has no sentiment. In online toy reviews, however, "recall" may more frequently be used in the sense of "withdraw from the market", as in "This toy should be *recalled*". Finally, even highly advanced sentiment analysis may be imprecise and subject to many false positives when used to identify safety concerns, since safety concerns are extremely rare and consumers may express strong negative sentiment about non-safety-related concerns such as durability, instructions, price, size, color, materials, entertainment value, and other aspects of the toy.

Given these limitations in generic sentiment analysis, there is reason to believe that conventional methods would not be maximally effective in uncovering safety concerns in the toy industry, and a more targeted approach is necessary.

2.2. Online reviews

As the world becomes increasingly digital, online reviews are becoming more popular and relied upon by consumers. Online reviews provide a source of consumer feedback and provide more transparency about products than ever before. Retailers such as Amazon, Target, Walmart, and Toys "R" Us provide a platform for customers to share their product experiences with others through online feedback. In analyzing the effect of word of mouth on sales, Chevalier and Mayzlin studied consumer reviews of books on two sites: Amazon.com and BarnesandNoble.com [10]. The study suggests that customers rely on review-text more heavily than on review summary statistics for books. Duan, Gu, and Whinston [13] find that online reviews and word of mouth are influential in driving movie box office sales.

Online review sources such as Amazon.com provide a valuable platform for uncovering common user safety concerns for certain product categories using automated computation. Amazon.com contains a large dataset of online reviews in relation to major product categories, such as "Toys and Games", where over two million consumer reviews have been written. This vast trove of consumer intelligence represents a treasure-chest of potential product safety insights.

2.3. Text mining

Text mining is becoming an increasingly popular method for analyzing big data and drawing conclusions. Researchers have used text from various sources, including discussion forums, news articles, customer reviews, and media reports, to extract data and summarize results to support decision making. Text-mining provides a valuable method to analyze a large textual source of customer feedback and deliver decision makers with valuable information for business process improvement. Spangler and Kreulen [34] analyzed unstructured customer data to determine a systematic approach for identifying common customer concerns. Coussement and van den Poel [12] also used text-mining to analyze a large dataset of inbound emails in order to automatically distinguish complaints from non-complaints.

Although there have been a multitude of text-mining studies applied to subject matters such as financial market predictions and general consumer attitudes, few methods have been developed to utilize text-mining in specifically targeting product safety issues. Abrahams et al. [3–5] applied text-mining to uncover safety defects in the automotive industry and provided a framework for applying this method to other product industries. Our study adapts this process to the children's toy industry.

2.4. Web and social media surveillance for public safety

Online news sources on the web have been used for surveillance of infectious disease outbreaks [45]—a flagship application of web mining for public safety purposes. The rise of social media on the web has sparked researchers to attempt to extract quantifiable data from this new prevalent form of communication. Social media sources include discussion forums, listservs, wikis, online communities such as social networks, usenet groups, customer product reviews, visitor comments, user-contributed new articles, and more. These sources may be used to

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

conduct information mining. For example, prior research has used text mining of online postings by automotive enthusiasts to uncover defects in motor vehicles [3–5]. There is strong evidence therefore that consumer reviews may be useful to uncover safety issues in other industries, such as the children's toy industry.

2.5. Summary

Past research regarding sentiment analysis, online reviews, text mining, and social media surveillance have all provided valuable techniques and information which have paved the way for automated toy safety surveillance from online reviews. This paper adapts and improves prior methods, which have not addressed safety surveillance in the children's toy industry.

3. Research questions and contributions

In this paper, we tackle three major research questions. Firstly, do online reviews in the toy industry contain substantial content related to injury existence and criticality? Secondly, when analyzing the content of the online reviews, can conventional sentiment analysis and other sentiment methods be used to distinguish safety concerns from reviews that do not mention safety concerns? If not, are there other characteristics that differentiate reviews that mention safety concerns from other reviews? Lastly, what alternative data sources and processing methods are available for smoke word discovery and injury severity scoring, and how do they compare in performance?

We make three major contributions in this paper. This is the first large-scale case study, to our knowledge, that confirms the usefulness of online reviews for safety surveillance in the toy industry. Secondly, we demonstrate that conventional sentiment analysis—though successfully applied previously to complaint detection in retail, finance, film, and other industries—must be adapted for safety concern detection and prioritization in the toy industry. Thirdly, we define a new class of toy "smoke" words that are valuable to the toy industry for this task and we describe a new procedure that provides robust safety concern discovery from online reviews, across multiple toy product categories and brands.

4. Methodology

In order to better understand the reporting of toy safety concerns in online reviews, we undertook a large empirical study of product reviews from the toy industry, specifically using the case study method. The case study method of theory building is widely accepted [7,14,20,28,44]. We adopted a research design consistent with earlier studies of consumer postings [17], and adhering to the guidelines of content analysis research [29].

4.1. Data sampling

For the *construction* of smoke lists—that is, lists of words likely to be indicative of safety concerns—we used two major data sources:

Toy-related hospitalizations (*CPSC NEISS*): Firstly, we used the National Electronic Injury Surveillance System (NEISS), years 2009–2014, from the United States Consumer Product Safety Commission's (U.S. CPSC) website. We compiled reports from the "Narrative" field and filtered by 38 toy product categories. We focused singularly on hospital admissions, or those narratives with disposition code "4", and arrived at a list of 587 toy-related injury narratives.

Toy-related recalls (*CPSC Recalls*): Secondly, we used CPSC Recall reports, years 1973–2015, from the United States Consumer Product Safety Commission's (U.S. CPSC) website. We filtered by 21 toy product categories and arrived at a list of 1065 toy-related recall reports (narratives).

To *test* the performance of the smoke word lists, we used the following data source:

Amazon.com toy reviews: We obtained online reviews from Amazon.com for the years 1999–2014 [27]. Of the total 146 million Amazon reviews, we found 2,234,519 in the category "Toys and Games". We randomly sampled 1.05 million Amazon reviews in the category "Toys and Games" for use as a test set.

4.2. Data processing

To *construct* smoke word lists, we proceeded as follows:

We computed the correlation coefficient (CC) [15] for each word in the *NEISS narratives set*, relative to a dummy document containing a single word not in the NEISS document set, to develop a ranking of prevalent unigrams. As the correlation coefficient is a document-based metric of term prevalence, a dummy document is necessary to ensure the denominator is non-zero. We manually filtered the resulting list, by excluding non-relevant terms that appeared in the scored list. Words excluded from the final smoke list included product words (e.g. vehicle, doll, helicopter), common industry words (e.g. toy, play, children), common English words (e.g. a, on, in, with, at), and common body parts (e.g. arm, leg, hair). Common product words and industry words were excluded in order to enhance generalizability to future product discussions. Common body parts were excluded because it was found that a smoke word list including these words resulted in a disproportionally high number of reviews associated with dolls and action figures. There were 110 remaining smoke words with a CC score greater than or equal to our chosen cutoff threshold (1.73). This threshold was chosen as the point at which word safety-relatedness appeared to noticeably diminish. The top 20 words in this smoke list are shown in the "Top NEISS smoke list" column in Table 1.

Next, we computed the correlation coefficient for each word in the *Recall narratives set*, relative to a dummy document containing a single word not in the Recall document set to develop a separate ranking of prevalent unigrams relative to Recall narratives. We again manually filtered by excluding non-relevant terms that appeared in the scored list. Words excluded from the final smoke list include product words, common industry words, common English words, and company names and trademarks (e.g. Fisher-Price, Playskool, Walmart). Company names were excluded to mitigate popularity bias: companies who sell more toys appear more frequently in recall announcements. This step is necessary to preserve generality, allowing the recognition techniques to be effective even as toy and retailer popularity vary over time. We retained 96 remaining recall smoke words with a CC score greater than or equal to our chosen cutoff threshold (1.73). The top 20 words in this smoke list are shown in the "Top recall smoke list" column in Table 1.

Twelve (12) words from the CPSC NEISS Smoke Word list overlapped with the CPSC recall smoke word list, as shown in the "Overlapping words" column in Table 1. This table shows minimal overlap between the two smoke word lists. A summary of the total smoke words used in each particular list, as well as the number of overlapping words, are shown below the word lists in Table 1.

For the purposes of comparison to conventional sentiment approaches, Table 1 also indicates, with the superscripts "*AFINN*" and "*GI*", which words appear also in the popular AFINN [31] and Harvard General Inquirer [23] dictionaries of words with negative sentiments. The final two summary rows in Table 1 indicate the total number of words in our full smoke lists which overlap with the full AFINN and Harvard General Inquirer negative words lists, and demonstrates negligible overlap.

As the toy product categories were not consistent between the NEISS and Recall narrative sets, we consolidated as appropriate to the 20 product categories shown in Table 2. Table 2 provides descriptive statistics

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

Table 1
Comparison of smoke word lists.

Top 20 NEISS smoke words		Top 20 Recall smoke words		Overlapping smoke words	
Word	CC Score	Word	CC Score	Word	
Fell	19.15	Choking[AFINN]	32.78	Swallowed	
Hit[GI]	9.27	Recalled	18.75	Ingested	
Swallowed	8.99	Recalls	16.82	Injury[AFINN,GI]	
Tripped	8.17	Lead	16.78	Ingestion	
Fracture[GI]	7.39	Hazard[GI]	14.73	Fall	
Femur	6.98	Recall	12.83	Laceration	
Fractured	6.37	Laceration	10.50	Choking[AFINN]	
Pain[AFINN,GI]	6.01	Burn[GI]	8.72	Aspiration	
Skull	5.72	Paint	8.34	Vomiting	
Admitted[AFINN]	5.62	Violation[GI]	8.14	Fire[AFINN,GI]	
Ingested	5.52	Fire[AFINN,GI]	8.08	Pieces	
Injury[AFINN,GI]	5.22	Hazards	7.53	Removed	
Hitting	4.67	Strangulation	6.95		
Landed	4.31	Ingestion	6.48		
Admit[AFINN]	4.31	Wooden	6.23		
Ingestion	4.06	Aspiration	6.15		
Fall	4.06	Posing	5.71		
Laceration	3.92	Injury[AFINN,GI]	5.03		
Humerus	3.65	Injuries	4.93		
Stuck[AFINN]	3.50	Internal	4.82		
# of words	110	# of words	96	# of words	12
# of AFINN negative	19	# of AFINN negative	16	# of AFINN negative	3
# of Harvard negative	12	# of Harvard negative	17	# of Harvard negative	2

for each toy category, including the total narratives for that toy category. For the narratives in each toy category, Table 2 shows averages for: word count, AFINN negative score, Harvard General Inquirer negative score, and smoke word count.

To test our smoke word lists, we ran three experiments, described below (Sections 4.2.1–4.2.3).

4.2.1. Data processing: Experiment 1

In our pilot study (Experiment 1), for the smoke word lists above, we used each smoke word list (NEISS, Recall) to score the large random sample of over one million Amazon.com toy reviews, incrementing the review's total accumulated score by the CC score for that word, each time the smoke word appeared in the review. We sorted the reviews from highest to lowest scoring. For each smoke list, we then ranked:

- The top 100 reviews, by summed CC score (using that smoke list metric)
- The bottom 100 reviews, by summed CC score (using that smoke list metric)

In the case of tied scores (e.g. bottom 100 reviews often had summed CC scores of zero (0), if no smoke words appeared in those reviews), we chose a random selection of reviews that had a tied score, to reduce bias. We then randomly mixed the top and bottom 100 reviews in each approach and hid the smoke scores (summed CC scores) to prevent bias in tagging each review. The lead member of the research team then manually tagged these reviews. In total, 400 reviews were manually tagged: 200 reviews derived from the NEISS smoke list scoring approach and 200 reviews derived from the Recall smoke list scoring approach. The tagging results are detailed in Section 5.1 below.

Table 2
Summary of scored narratives by product sub-category.

Product category	NEISS narratives					Recall narratives				
	Narrative count	Mean word count	Mean AFINN negative	Mean GI negative	Mean NEISS smoke count	Narrative count	Mean word count	Mean AFINN negative	Mean GI negative	Mean recall smoke count
Building sets	53	15.7	0.9	0.8	2.1	24	298.3	13.2	8.6	10.8
Child arts and crafts, crayons and chalk	18	16.3	0.9	0.8	1.7	7	408.3	11.1	10.9	9.4
Costume/children's jewelry	–	–	–	–	–	102	304.5	7.8	5.8	11.4
Dolls, plush toys, and action figures	18	17.3	1.3	1.1	2.2	246	301.6	8.1	6.3	10.8
Marbles	19	11.3	0.7	0.5	2.1	–	–	–	–	–
Playground items	–	–	–	–	–	16	334.1	11.3	8.8	10.3
Puzzles	–	–	–	–	–	20	340.0	9.5	8.6	13.0
Toy balls	85	19.1	0.8	1.2	2.7	48	316.4	8.4	7.4	11.0
Toy chests/trunks	–	–	–	–	–	17	310.4	13.4	6.5	11.5
Toy infants/cribs/strollers	–	–	–	–	–	32	329.0	9.0	8.0	11.0
Toy miscellaneous	173	15.9	1.0	0.9	2.4	160	290.0	7.9	6.8	11.3
Toy planes	–	–	–	–	–	14	316.1	9.1	8.2	11.8
Toy play sets/activity sets	3	14.0	2.0	0.7	3.3	83	324.2	9.7	7.1	12.3
Toy ride-on	122	17.0	0.8	0.7	2.5	40	292.8	10.0	9.6	10.8
Toy sports	9	21.4	1.7	1.1	2.0	12	329.6	7.3	6.6	9.2
Toy telephones	–	–	–	–	–	18	351.6	12.7	8.8	11.8
Toy vehicles	44	18.1	0.9	0.9	2.3	204	327.0	10.9	8.5	11.8
Toy weapons	15	20.1	1.6	1.7	2.3	11	251.6	10.4	8.2	8.3
Toys for bathtub	–	–	–	–	–	11	381.7	12.0	9.9	12.7
Wagons (children's)	28	15.5	0.4	0.6	3.0	–	–	–	–	–
Total	587	16.8	0.9	0.9	2.4	1065	311.6	9.2	7.3	11.3

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

4.2.2. Data processing: Experiment 2

In a follow-up study (Experiment 2), for the two smoke word lists above, we used each smoke word list (NEISS, Recall) to score the online reviews, incrementing the review's total accumulated score by the CC score for that word, each time the smoke word appeared in the review. We sorted the reviews from highest to lowest scoring then, for each smoke list, and ranked:

- The top 400 reviews, by summed CC score (using that smoke list metric)
- The bottom 400 reviews, by summed CC score (using that smoke list metric)

We randomly selected the bottom 400 reviews with a score of 0, since thousands of reviews had zero scores. We arrived at a total of 1600 reviews: 800 reviews derived from the NEISS smoke list scoring approach and 800 reviews derived from the Recall smoke list approach. However, in contrast to Experiment 1, we also used 800 completely random, unscored reviews to create a baseline for comparison. This resulted in a total of 2400 reviews in Experiment 2. We randomized these, to ensure taggers would not be biased by the order or co-occurrence of reviews. To reduce tagger bias further, nine different undergraduate students, all majoring in business information technology, tagged these 2400 reviews, following the protocol described below in Section 4.3.2, and tagging at least 400 reviews each.

We compared sentiment analysis vs. smoke word surveillance for Experiment 2 only, as it provided a more comprehensive and unbiased dataset than Experiment 1. We used the AFINN, ANEW, and Harvard General Inquirer (negative sentiment) dictionary metrics, as well as the SentiStrength, Opinion Finder Negative, and Amazon star rating sentiment methods to test how effective these methods were in identifying safety issues in the set of 1600 Recall- and NEISS smoke word list-scored reviews. We used a *t*-test to measure the difference in means in these sentiment scores and these smoke-scores between reviews that mentioned safety concerns and those that did not. Table 4 shows the results.

4.2.3. Data processing: Experiment 3

In another follow-up study (Experiment 3), we assessed the performance of two sentiment methods in scoring the random sample of 1.05 million Amazon toy reviews: (1) context-aware sentiment scoring, using SentiStrength, and (2) consumer-assigned sentiment, using overall Star Rating from the original review. For SentiStrength, we looked only at negative scores to ensure positive sentiment did not mask negative sentiment. We found:

- The most negative 400 reviews, by SentiStrength negative score
- The least negative 400 reviews, by SentiStrength negative score

For Amazon star rating analysis, we sorted the set of reviews using the review author's overall star rating from Amazon. We found:

- 400 one-star reviews
- 400 five-star reviews

In both cases above, where scores were tied, we randomly selected from tied scores.

4.3. Data coding

When deciding on categories for tagging reviews, we adapted our coding scheme from the CPSC "Class A–E product hazards" [39]. Since the definition of Class A through E product hazards is nuanced, we simplified these hazard categories to a subset that could be reliably coded by laypersons.

4.3.1. Data coding: Experiment 1

In our pilot study (Experiment 1), to determine whether the smoke word lists were effective, Amazon reviews scored using each smoke word list were tagged using the following tagging protocol:

- Injury existence: safety issue vs. non-issue
- Injury timing: actual injury occurred vs. potential injury
- Injury severity: actual minor injury vs. actual major injury

For each review, the tagger determined injury existence, or whether a specific injury or safety issue was explicitly mentioned in the customer review. If no injury was found in the review, the "Minor Injury Occurred", "Major Injury Occurred", and "Potential Safety Issue" fields were all tagged "No". An example of a review with no mention of a safety issue was:

> "Got it for my daughter for Christmas. Two people, takes less than an hour to assemble. Relatively simple instructions, and easy assembly (with a partner). Sturdy and enjoyable."

If there was an injury or safety concern explicitly mentioned in the review, the tagger assessed the injury timing. If the reviewer did not mention a specific injury that occurred but expressed concern about a potential safety issue with using the toy product, then only the "Potential Safety Issue" field was marked as "Yes". An example of a review with a potential safety issue was:

> "…However, understanding where this toy came from, I immediately became concerned for the use of lead paint. I couldn't find any warnings on the original boxes verifying that no lead paint was being used. With how these toys smelled, I would never want to let my nephews touch them and then put their hands in their mouths…"

Here, the reviewer doesn't mention an injury that actually occurred, but rather expresses concern that use of this product could result in a safety issue. If the reviewer did mention an actual injury that occurred as a result of using the product, then the tagger determined whether it was a minor injury or major injury. A minor injury was an incident that did not require a doctor's visit or hospitalization, such as a rash, minor cut, or red mark. An example of a review in which a minor injury occurred was:

> "…My other problem is the handle. It isn't solid on the bottom side. He holds this vacuum for hours at a time very tightly because he is excited, and the edges of the handle chafe his skin. We are going to have to wrap it in foam tape or something…"

If a minor injury occurred, the "Minor Injury Occurred" field was tagged "Yes", as well as the "Potential Safety Issue" field. A major injury was an incident that probably caused significant pain or concern, or ended with a doctor's visit or hospital admission, such as a choking incident, concussion, or deep cut. An example of a review in which a major injury was mentioned is:

> "…I liked these crayons a lot until my 2 year old found one of his older siblings crayons and decided to eat them. They truly are a choking hazard and because they are plastic, I wonder if they are more of one than wax crayons. My son had a piece of then removed from his lungs and throat and is currently on a ventilator…"

If a major injury was reported, all fields—"Minor Injury Occurred", "Major Injury Occurred", and "Potential Safety Issue" were tagged "Yes".

4.3.2. Data coding: Experiment 2

In our more expansive follow-up study (Experiment 2) to further validate whether the smoke word lists were effective, we tagged the selected reviews (Section 4.2.2) using the following tagging protocol:

- Defect existence: safety defect vs. performance defect vs. no defect

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

For each review, the tagger determined whether a specific injury or safety concern was mentioned in the customer review. If no injury or safety concern was found in the review, the tagger would determine if the review indicated a "Performance Defect" or "No Defect". Performance defects are not relevant for the current study on safety concerns, but may be used in future research.

In Experiment 2, the previous tagging categories of "Minor Injury Occurred", "Major Injury Occurred", and "Potential Future Injury" were combined and simply classified as "Safety Defect". We define a Safety Defect as a safety concern expressed by the review author as judged by a layperson review tagger. This may or may not be a bona fide safety issue verified by a consumer safety regulatory agency, such as CPSC. Example of reviews with safety defects can be found in Section 4.3.1, earlier.

Most reviews (1559 out of 2400) were tagged by multiple members of the tagging team. We used Cohen's kappa statistic (κ) [11] to determine inter-rater reliability. The procedure we used to calculate κ was as follows:

- Tagger A was the authority tagger (the lead member of the research team)
- Tagger B was the conservative combined opinion of all other taggers, meaning tagger B tagged as safety defect if any of the members of the group said the review was a safety defect. Though voting is commonly used to determine a final decision, the cost of a false negative for safety concerns is high, thus a conservative strategy is essential.

We observed $\kappa = 0.692$ ($n = 488$; 463 agreements; 25 disagreements; 94.9% agreement). $\kappa = 0.692$ is regarded as "substantial" agreement by Landis & Koch [24] and as "fair to good" agreement by Fleiss et al. [18]. To compute κ, we compared the authority tagger's tags to the conservative vote of the remaining taggers. The authority tagger tagged 488 reviews in Experiment 2, therefore κ was computed for every one of the authority tagger's reviews ($n = 488$) versus the combined opinion of the remainder of the tagging team on those 488 reviews. We mitigated the potential for over-estimation of κ by first having the non-authority taggers complete their tagging, and then having the authority tag a random selection of all reviews tagged by the non-authority taggers. The alternative method of having the non-authority taggers first tag authority-tagged reviews, and then allowing the non-authority taggers to continue tagging only if reliability is established, could have inflated κ, as it evaluates κ up-front, rather than over the full duration of non-authority tagging, and non-authority reliability could diminish with fatigue.

Additional review was performed by the lead author to make sure there were no false positives.

It can be observed that 841 reviews (2400–1559) were tagged by only a single person (meaning they were not tagged by multiple taggers). This does not mean that these 841 reviews were all tagged by the same person, but rather that there were 841 reviews where only one person tagged the review. Single-tagger reviews were included in the review set, as κ establishes that they were reliably tagged. The justification for this tagging procedure is that, once tagger reliability is established, it allows a larger number of items to be tagged, with limited human resources, as every review does not need to be tagged by multiple taggers. The limitation of this tagging procedure is that a single tagger's opinion is relied on for a subset of the reviews, meaning safety defects may be missed if that tagger was fatigued or lost accuracy. A trade-off therefore exists between tagging expense (the labor expense is higher if every review is manually tagged by multiple taggers) versus tagging assurance (you can be more assured of finding safety defects if multiple taggers look at every review).

The data coding approach of determining tagger reliability, and then including items tagged by only a single tagger once taggers are known to be reliable, is an established method in the field of Content Analysis [29] and qualitative research studies [21], and has been used in prior defect discovery studies [3,5].

4.3.3. Data coding: Experiment 3

Experiment 3 employed the same tagging protocol as Experiment 2. Reviews were again randomized to ensure taggers would not be biased by the order or co-occurrence of reviews. To ensure consistency, four of five taggers used in Experiment 3 were members from the original Experiment 2 team. Each of the taggers coded at least 400 reviews. The reliability of the authority tagger, who tagged 400 reviews, was cross-checked against the combined opinion of all other taggers, in the same manner as for Experiment 2 above. We observed $\kappa = 0.453$ ($n = 400$; 393 agreements; 7 disagreements; 98.3% agreement). $\kappa = 0.453$ is regarded as "moderate" agreement by Landis and Koch [24] and as "fair to good" agreement by Fleiss et al. [18]. The lower κ score in Experiment 3 is to be expected, as safety concerns were less frequently identified by the sentiment analysis techniques in Experiment 3, and Cohen's κ is known to be understated when the two classes being coded are not equiprobable.

5. Results and evaluation

In this section, we evaluate the performance of our industry-specific smoke word lists for safety concern discovery, and compare results to traditional sentiment analysis approaches.

5.1. Experiment 1 results

Sixteen minor, major, or potential future injury mentions were found in the top 100 reviews scored using the NEISS smoke word list, whereas only 1 minor injury was found in the bottom 100 reviews scored using the NEISS smoke word list.

A total of 42 minor, major, or potential future injury mentions were found in the top 100 reviews scored using the Recall smoke word list whereas there were no mentions of injury concerns in the bottom 100 reviews scored using the Recall smoke word list. Table 3 gives an overview of these findings.

5.2. Experiment 2 results

Eleven safety defects or concerns were found in the "baseline" set of 800 random "Toys and Games" reviews (i.e. 5.5 safety defects expected per 400 reviews).

Using the NEISS smoke word list to score reviews, 44 out of the top 400 scored reviews mentioned safety concerns while only 3 out of the bottom 400 mentioned safety concerns.

Using the Recall smoke word list to score reviews, 155 out of the top 400 scored reviews mentioned safety concerns, while only 2 safety concerns were mentioned in the bottom 400 scoring reviews.

The top half of Table 4 provides a summary of these findings, comparing smoke list-scored reviews to baseline random reviews to determine effectiveness. The Recall smoke list proved to be the most effective discovery method, as it uncovered the most safety concerns.

We used the chi-squared test to determine whether there were significantly more defects in the various subsets of reviews we compared. We found that the top 400 NEISS scored reviews revealed significantly more ($p < 0.001$) safety concerns than the bottom 400 NEISS scored reviews and 400 baseline reviews. We also found that the top 400 Recall scored reviews revealed significantly more ($p < 0.001$) safety concerns

Table 3
Experiment 1 tagging results.

Review set		Minor injury mentioned	Major injury mentioned	Potential safety issue mentioned	Performance defect or non-defect
NEISS	Top 100	5	1	16	84
	Bottom 100	1	0	0	99
Recall	Top 100	2	1	42	58
	Bottom 100	0	0	0	100

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

Table 4
Safety concerns found in top- and bottom-ranked reviews scored using NEISS and recall smoke lists (Experiment 2).

Review set		Safety concerns	No safety concerns	Chi-square *p*-value
Smoke: NEISS	Top 400	44	356	<0.001**
	Bottom 400	3	397	0.28
Smoke: Recall	Top 400	156	244	<0.001**
	Bottom 400	2	398	0.13
	Sub-total	205	1395	
	Total	1600		
Baseline (out of 400)		*5.5*	*394.5*	

Scoring metric	Safety concerns mean score	No safety concerns mean score	*t*-test *p*-value
Smoke: NEISS	18.62	15.07	<0.001**
Smoke: Recall	83.05	21.48	<0.001**
AFINN	3.49	10.65	<0.001**
ANEW	168.80	175.37	0.43
GI Negative	6.63	6.05	0.09
SentiStrength Negative	−2.87	−2.17	<0.001**
OpinionFinder Negative	6.20	4.55	0.01**
Amazon Star Rating	2.36	3.83	<0.001**

**, Indicates strong statistical significance at the 99% confidence level.

Table 5
AFINN words contributing most negativity in safety concerns in validation set.

Word	Total contribution to AFINN negative score across all safety concerns	% of Total negativity expressed in safety concerns	Cumulative % of total negativity expressed in safety concerns
Choking	−466	19%	19%
Warning	−111	5%	24%
No	−86	4%	27%
Bad	−54	2%	30%
Loose	−54	2%	32%
Problem	−42	2%	34%
Blocks	−40	2%	35%
Poor	−38	2%	37%
Risk	−38	2%	38%
Hard	−36	1%	40%
Lost	−36	1%	41%
Damage	−33	1%	43%
Worry	−33	1%	44%
Stuck	−30	1%	45%
Disappointed	−28	1%	46%
Hurt	−28	1%	48%
Broke	−25	1%	49%
Die	−24	1%	50%
Horrible	−24	1%	51%
Worse	−24	1%	52%

than the bottom 400 Recall scored reviews, 400 baseline reviews, and top 400 scored NEISS reviews. Both smoke lists were effective in uncovering safety concerns in children's toys, while the Recall list was more effective than the NEISS list.

Focusing only on the 1600 reviews scored using either the NEISS smoke list or the Recall smoke list in Experiment 2 ("the validation set"), we used the AFINN [31], ANEW [9], Harvard General Inquirer [23], SentiStrength [37,38] and OpinionFinder [42,43] methods, and Amazon overall Star Ratings to test whether conventional sentiment analysis methods produce significantly different scores for the safety concerns we identified vs. non-safety concerns in the validation set. For Harvard GI, SentiStrength, and OpinionFinder, we used only negative sentiment, to avoid the masking effect of positive sentiment. We used the standard Student's *t*-test to the measure the difference in means in these sentiment scores between reviews that contained safety concerns and reviews that did not mention safety concerns. The bottom half of Table 4 summarizes our results.

Our *t*-tests indicate that both the NEISS and Recall smoke list approaches yielded strongly statistically significant score differences ($p < 0.001$) between reviews with safety concerns and those without safety concerns in the validation set. The ANEW and Harvard General Inquirer (negative sentiment) dictionary metrics did not exhibit a statistically significant difference between reviews marked as safety concerns and those marked as no safety concerns in the validation set (p-values = 0.43 and 0.09, respectively). In contrast, our *t*-tests indicate with strong statistical significance ($p \le 0.01$), that reviews with safety concerns have significantly more negative sentiment, when scored with all other sentiment techniques (AFINN, SentiStrength negative, OpinionFinder negative, and overall Amazon Star Rating).

We noticed that the average AFINN scores were still positive in reviews associated with safety concerns. This was an unexpected finding, since we expected safety concerns to have a *strongly negative* AFINN score. Of the 1598 negative valence words in the AFINN dictionary (2477 total AFINN words), only 250 words contributed to negative sentiment scores of the reviews tagged as safety concerns in our validation set. The top 6 negative AFINN words accounted for one third of total negativity expressed, the top 18 AFINN words accounted for one half of all negativity expressed, and the top 80 AFINN words accounted for 80% of all negativity expressed. Table 5 shows the top 20 AFINN words in reviews tagged as safety defects in the validation set, and their contribution to total negativity across all reviews tagged as safety defects.

For each toy sub-category in Experiment 2, we computed a Category Hazard Rating (CHR), which we define as the ratio of safety defects found in a toy sub-category (i.e. True Positives), relative to the number of safety defects expected to be present in that sub-category (denoted "E(P)"), given the proportion of reviews in that sub-category, and given the baseline rate of safety defects across all categories. The CHR for each toy sub-category is shown in column 9 in Table 6. High CHRs indicate sub-categories with a disproportionally high number of safety defects, in the validation set. Note that all sub-categories display a disproportionally high number of safety defects, as the NEISS and Recall smoke list techniques are highly effective at discovering safety defects across all sub-categories. Table 6 indicates that the sub-categories Games and Baby and Toddler Toys are of most concern with regard to number of hazards, and the sub-categories Party Supplies and Stuffed Animals and Plush Toys are of least concern with regard to the number of hazards. Caution in interpreting these numbers should, however, be exercised, as hazards have different immediacy and severity, and a small number of severe hazards can still be of high concern.

Table 6 shows count of true-positives (TP), false-positives (FP), true-negatives (TN), and false-negatives (FN), as well as Precision, and Recall, for safety-defect discovery in each major Amazon toy product sub-category.

For *Precision*, note that, given the low baseline rate of safety defects (11 safety defects per 800 = 1.4%) in our random sample of reviews, the minimum acceptable precision necessary for a scoring method to achieve statistical significance at the 95% confidence level is only 2.25% (18 safety defects per 800; using Chi-test). This acceptable precision threshold is modest, and has easily been achieved across all toy sub-categories. As shown later, in Experiment 3 (Table 7), the best benchmark (i.e. SentiStrength) has 5.25% (=21/400) precision. It is widely established in machine learning [8,16], medical diagnostics [22], and direct marketing [25], that classification precision should be considered relative to the baseline rate of occurrence of the response class in a random sample: a low precision classifier is still highly beneficial when precision is substantially greater than the baseline rate. Consider that, even in the sub-category with the lowest precision in Table 6 (9% precision in the Hobbies category), we find six times (= 9%/1.4%) more safety defects in the subset of reviews predicted to have safety defects by the scoring method, than would be expected in a random selection of reviews (1.4%).

Nevertheless, the comparatively low precision in categories such as Hobbies, Arts and Crafts, and Action Figures and Statues, when contrasted to other sub-categories, indicates that sub-category-specific smoke word lists may be beneficial for further improving safety defect

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

Table 6
Reviews, safety defects, and hazard ratings by product sub-category.

Product sub-category	Total reviews	Percent of reviews	E(P)	TP	FP	TN	FN	CHR = TP/E(P)	Precision = TP/(TP + FP)	Recall = TP/(TP + FN)
Action figures and statues	220,100	9.8%	1.1	14	95	75	0	12.92**	14/109 = 13%	14/14 = 100%
Arts and crafts	104,517	4.7%	0.5	5	41	44	0	9.72**	5/46 = 11%	5/5 = 100%
Baby and toddler toys	61,188	2.7%	0.3	31	58	83	0	102.92**	31/89 = 35%	31/31 = 100%
Building toys	102,464	4.6%	0.5	12	25	42	0	23.79**	12/37 = 32%	12/12 = 100%
Dolls and accessories	149,171	6.7%	0.7	12	49	52	0	16.34**	12/51 = 20%	12/12 = 100%
Dress up and pretend play	96,781	4.3%	0.5	27	35	35	0	56.67**	27/62 = 44%	27/27 = 100%
Electronics for kids	112,315	5.0%	0.6	8	41	55	0	14.47**	8/49 = 16%	8/8 = 100%
Games	311,987	14.0%	1.5	162	502	658	5	105.48**	162/668 = 24%	162/167 = 97%
Grown-up toys	49,518	2.2%	0.2	2	13	18	0	8.20*	2/15 = 13%	2/2 = 100%
Hobbies	116,157	5.2%	0.6	5	49	38	0	8.74**	5/54 = 9%	5/5 = 100%
Learning and education	128,765	5.8%	0.6	29	41	51	1	45.75**	29/70 = 41%	29/30 = 97%
Novelty and gag toys	83,505	3.7%	0.4	8	14	22	0	19.46**	8/22 = 36%	8/8 = 100%
Party supplies	69,040	3.1%	0.3	2	7	39	0	5.88**	2/9 = 22%	2/2 = 100%
Puzzles	76,727	3.4%	0.4	4	22	22	0	10.59**	4/26 = 15%	4/4 = 100%
Sports and outdoor play	139,978	6.3%	0.7	14	33	56	1	20.32**	14/47 = 30%	14/15 = 93%
Stuffed animals and plush	191,871	8.6%	0.9	4	16	63	0	4.23**	4/20 = 20%	4/4 = 100%
Remote control and play vehicles	131,377	5.9%	0.6	20	52	53	1	30.92**	20/72 = 20%	20/21 = 95%
Tricycles	40,376	1.8%	0.2	3	14	12	0	15.09**	3/17 = 18%	3/3 = 100%
Total	2,234,519			362	1107	1418	8			

**, Indicates statistical significance at the 99% confidence level.
E(P) is the expected safety defects in this category per 800 reviews, since there were 800 reviews predicted to be safety concerns (Top 400 NEISS + Top 400 Recall).
E(P) = Percent of reviews × 800 × baseline rate of safety defects found in random sample of full data set (i.e. 1.4%).
Some items fall into multiple sub-categories as Amazon often classifies toys in multiple sub-categories. Only major Amazon toy sub-categories are shown; minor sub-categories are omitted for brevity. Regarding the correspondence between Tables 4 and 6: note that 800 reviews are predicted to be safety concerns (Top 400 NEISS + Top 400 Recall) according to Table 4. However, the sum of Total TP plus Total FP in Table 6 sums to more than 800, as reviews may be counted under multiple sub-categories.

discovery. For example, in "Action Figures and Statues" we found that body part smoke words can result in false positives, as the review writer is mentioning a body part of the toy, not a body part of an injured person. Further research is thus needed to refine category-specific smoke word lists.

For *Recall*, note that this metric should be interpreted cautiously, as Table 6 captures only the small, biased sample of toy reviews (namely, the top scoring reviews, by smoke score) that were manually tagged. Without tagging the full toy dataset, we cannot ascertain the true number of false negatives (safety defects missed), though, as shown above, we estimate (with 95% confidence) that safety defects occur in up to 2.25% of reviews.

5.3. Experiment 3 results

Table 7 shows that, using SentiStrength negative scores to score reviews, 21 out of the top 400 scored (most negative) reviews mentioned safety concerns (5.25%), while only 4 out of the bottom 400 (least negative) mentioned safety concerns (1%). A chi-squared test comparing safety concerns in the 400 most negative SentiStrength reviews to the baseline rate of 5.5 safety concerns per 400 random reviews (1.4%) indicates there are a significantly larger proportion of safety defects in the most negative SentiStrength reviews, versus a random sample of reviews ($p < 0.001$).

Table 7
Safety concerns found in top- and bottom-ranked reviews scored using SentiStrength and overall Amazon Star Rating (Experiment 3).

Review set		Safety concerns	No safety concerns	Chi-square *p*-value
SentiStrength	400 Most Negative	21	379	<0.001**
	400 Least Negative	4	396	0.52
Star Rating	400 Random One Star	6	394	0.83
	400 Random Five Star	0	400	0.02*
Baseline (out of 400)		*5.5*	*394.5*	

*, Indicates statistical significance at the 95% confidence level.
**, Indicates strong statistical significance at the 99% confidence level.

Using Amazon overall Star Ratings to sort reviews, 6 out of 400 random 1-star reviews mentioned safety concerns (1.5%), while 0 safety concerns were mentioned in 400 random 5-star reviews (0%). Five-star reviews mention statistically significantly fewer safety concerns than the baseline (p-value = 0.02). A chi-squared test comparing 1-star reviews to the same baseline rate indicates no statistically significant difference between the proportion of safety concerns in 1-star reviews, and the proportion of safety concerns in a random sample (p-value = 0.83).

Table 7 provides a summary of these findings, comparing these results to the baseline set of random reviews, to determine effectiveness. The SentiStrength method in Experiment 3 proved to be the most effective safety concern discovery method, as it uncovered the most safety concerns. However, this method is still not as effective as the NEISS and Recall smoke lists (Experiment 2, earlier).

5.4. Summary of results

Smoke word approaches are highly effective in identifying toy reviews that mention safety concerns. Both Experiment 1 and Experiment 2 provide evidence for the effectiveness of the smoke word approaches. Sentiment methods, although statistically significant in differentiating reviews, did not offer nearly as much disparity in safety concerns vs. no safety concerns as the NEISS and Recall smoke lists (Experiment 2).

Experiment 3 showed that, although SentiStrength scores were effective in finding reviews with safety concerns, they were not nearly as effective as the NEISS and Recall smoke lists in Experiment 2. While there is a strongly statistically significant difference in sentiment between safety-defects and non-safety-defects in many sentiment analysis methods, sentiment analysis is not the most effective means of discovering safety defects. In the top 400 scoring reviews of each approach, the Recall smoke word method finds significantly more ($p < 0.001$) safety concerns than the SentiStrength method (156 vs. 21). Sentiment analysis appears to perform far less effectively because safety defects are rare and are buried in mounds of irate postings containing non-safety related complaints about the toys, such as poor durability, difficult assembly, and others. Sentiment analysis is therefore insufficiently specific at identifying safety concerns, and precision can be significantly improved using smoke word approaches.

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

6. Limitations and future work

Our study has some limitations that we would like to address in future research. Firstly, only unigrams were employed in developing smoke term lists from the CPSC NEISS and Recall narratives. Secondly, while data from hundreds of brands and models was used, data from only a single retailer was used (Amazon). Additionally, Amazon's "Toys and Games" category included a small portion of reviews that weren't relevant to children's toys (e.g. adult games and toys). Finally, we did not use the available disambiguation feature of the Harvard General Inquirer software to disambiguate word senses.

In future work, we will look to expand beyond unigrams and include bigrams and trigrams in our smoke lists to score reviews. An example of a bigram term likely to be effective would be "choking hazard" whereas "fell out of" could be an effective trigram term. We will look at another smoke word generation approach involving contrasting narratives (e.g. Recall and NEISS) with an equal number of random Amazon reviews to determine which terms are particularly prevalent in smoke term sources. We plan to create separate dictionaries for major childhood injury categories (trips and falls, choking, drowning, etc.), to specifically identify particular hazard types. Future work should include stemming the smoke lists and the text to be scored (to identify root word forms) and perhaps refining the smoke lists more. Scored reviews may be alternatively ranked by adjusting for overall review length: computing, for example, the percentage of smoke words, or smoke phrases per 100 words in the review. Future tagging may encompass other classifications besides safety issues, such as performance defects, and may include a larger scored data set, a higher volume of manually tagged reviews, and more reviews tagged by multiple taggers.

7. Implications for practice and research

Practitioners can use our research methods to more rapidly identify children's toys with safety concerns. These methods should garner particular interest from the U.S. CPSC, as thousands of new toys are introduced each year, with limited staff resources to fully test and evaluate every toy that reaches store shelves. Furthermore, retailers could monitor toys more extensively using our approach in order to limit recalls that would otherwise harm their reputation.

Our research could lead to possible action items for retailers and regulatory agencies. For example, manufacturers and retailers selling more than a threshold sales volume (e.g. 10,000 units a week) could be mandated to have a certified online review and email surveillance process. Our process for scoring online reviews could be used as a basis for an international, standard safety surveillance certification for the toy industry, in English-speaking countries. The process would need to be adapted for other languages.

Researchers could adapt our methods to apply to safety-defect and performance-defect discovery in other industries, such as kitchen and home appliances, furniture, electronics, cosmetics, food, and other consumer products. Tagging classifications could be adapted as appropriate.

8. Summary and conclusions

In this paper we evaluated a text mining approach for discovering safety concerns mentioned in children's toy reviews. We adapted prior defect discovery systems [3–5] to the children's toy industry. We used public U.S. CPSC records to develop two different "smoke lists": one from CPSC National Electronic Injury Surveillance System (NEISS) narratives and the other from CPSC Recall reports. We used these smoke lists to score over one million Amazon reviews under the category "Toys and Games". We conducted three experiments to determine the effectiveness of the smoke list approaches, and contrast to sentiment approaches. We determined that this customized approach was indeed effective, using both chi-squared and *t*-tests of statistical significance.

Our findings highlight that practitioners and researchers need to be cautious in applying generic sentiment analysis to the discovery of safety concerns in online reviews, since these conventional tools appear to be only moderately effective. In contrast, the methods and toy industry-specific smoke words outlined in this paper appear to show strong promise for monitoring children's toys for safety concerns, and could potentially be adapted to other industries in the future.

References

[1] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Transactions on Information Systems (TOIS) 26 (3) (2008) 12.
[2] V.M. Abraham, C.E. Gaw, T. Chounthirath, G.A. Smith, Toy-related injuries among children treated in US emergency departments, 1990–2011, Clinical Pediatrics 54 (2) (2014) 127–137.
[3] A.S. Abrahams, W. Fan, G.A. Wang, Z.J. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery, Production and Operations Management 24 (6) (2015) 975–990.
[4] A.S. Abrahams, J. Jiao, W. Fan, G.A. Wang, Z. Zhang, What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings, Decision Support Systems 55 (4) (2013) 871–882.
[5] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decision Support Systems 54 (1) (2012) 87–97.
[6] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, The Journal of Finance 59 (3) (2004) 1259–1294.
[7] I. Benbasat, D.K. Goldstein, M. Mead, The case research strategy in studies of information systems, MIS Quarterly 11 (3) (1987) 369–386.
[8] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.
[9] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings, Technical Report C-1, the Center for Research in Psychophysiology, University of Florida, 1999.
[10] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: online book reviews, Journal of Marketing Research 43 (3) (2006) 345–354.
[11] J. Cohen, Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit, Psychological Bulletin 70 (4) (1968) 213.
[12] K. Coussement, D. Van den Poel, Improving customer complaint management by automatic email classification using linguistic style features as predictors, Decision Support Systems 44 (4) (2008) 870–882.
[13] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter?—An empirical investigation of panel data, Decision Support Systems 45 (4) (2008) 1007–1016.
[14] K.M. Eisenhardt, Building theories from case study research, Academy of Management Review 14 (4) (1989) 532–550.
[15] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decision Support Systems 40 (2) (2005) 213–233.
[16] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.
[17] B.J. Finch, Internet discussions as a source for consumer product customer involvement and quality information: an exploratory study, Journal of Operations Management 17 (5) (1999) 535–556.
[18] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, John Wiley & Sons, 2013.
[19] Defective toys: definitely not child's play, Nov. 22, 2015 (Retrieved from) https://www.justice.org/sections/newsletters/articles/defective-toys-definitely-not-childs-play.
[20] B.G. Glaser, A.L. Strauss, The Discovery of Grounded Theory: Strategies for Qualitative Research, Transaction Publishers, 2009.
[21] K.A. Hallgren, Computing inter-rater reliability for observational data: an overview and tutorial, Tutorials in Quantitative Methods for Psychology 8 (1) (2012) 23.
[22] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
[23] E.F. Kelly, P.J. Stone, Computer Recognition of English Word Senses, North-Holland, 1975.
[24] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174.
[25] C.X. Ling, C. Li, Data Mining for Direct Marketing: Problems and Solutions, KDD, 1998 73–79.
[26] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, The Journal of Finance 66 (1) (2011) 35–65.
[27] J. McAuley, R. Pandey, J. Leskovec, Inferring Networks of Substitutable and Complementary Products, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM 2015, pp. 785–794.
[28] D.M. McCutcheon, J.R. Meredith, Conducting case study research in operations management, Journal of Operations Management 11 (3) (1993) 239–256.
[29] K.A. Neuendorf, The Content Analysis Guidebook, Sage, 2002.
[30] Wal-Mart Recalls 174,000 Dolls Over Burn Risk, Nov. 22, 2015 (Retrieved from) http://www.nbcnews.com/business/consumer/wal-mart-recalls-174-000-dolls-over-burn-risk-n62071.
[31] F.Å. Nielsen, A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs, Proceedings of the 1st Workshop on Making Sense of Microposts, Heraklion, Crete 2011, pp. 93–98.

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016

[32] C. Oh, O. Sheng, Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement, Proceedings of the 32nd International Conference on Information Systems, Shanghai, China, Paper, 17 2011, pp. 1–19.
[33] Annual Sales Data, US Domestic Markets, The NPD Group, Nov. 22, 2015 (Retrieved from http://www.toyassociation.org/TIA/Industry_Facts/salesdata/IndustryFacts/Sales_Data/Sales_Data.aspx?hkey=6381a73a-ce46-4caf-8bc1-72b99567df1e-VhlRQPl3ldh).
[34] S. Spangler, J. Kreulen, Mining the Talk: Unlocking the Business Value in Unstructured Information, IBM Press, 2007.
[35] L. Story, D. Barboza, Mattel recalls 19 million toys sent from China, The New York Times, Aug. 15, 2007 The URL is http://www.nytimes.com/2007/08/15/business/worldbusiness/15imports.html?_r=0.
[36] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: quantifying language to measure firms' fundamentals, The Journal of Finance 63 (3) (2008) 1437–1467.
[37] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, Journal of the American Society for Information Science and Technology 63 (1) (2012) 163–173.
[38] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, Journal of the American Society for Information Science and Technology 61 (12) (2010) 2544–2558.
[39] Recall Handbook, United States Consumer Product Safety Commission, Nov. 22, 2015 (Retrieved from http://www.cpsc.gov/PageFiles/106141/8002.pdf).
[40] Toy-Related Deaths and Injury Calendar, United States Consumer Product Safety Commission, Nov. 22, 2015 (Retrieved from http://www.cpsc.gov/Global/Research-and-Statistics/Injury-Statistics/Toys/ToyReport2013.pdf).
[41] Toy Recall Statistics, United States Consumer Product Safety Commission, Nov. 22, 2015 (Retrieved from http://www.cpsc.gov/en/Safety-Education/Toy-Recall-Statistics/).
[42] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, OpinionFinder: a System for Subjectivity Analysis, Proceedings of Hlt/Emnlp on Interactive Demonstrations, Association for Computational Linguistics 2005, pp. 34–35.
[43] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics 2005, pp. 347–354.
[44] R.K. Yin, Case Study Research: Design and Methods, Sage Publications, 2013.
[45] Y. Zhang, Y. Dang, H. Chen, M. Thurmond, C. Larson, Automatic online news monitoring and classification for syndromic surveillance, Decision Support Systems 47 (4) (2009) 508–517.

Matt Winkler holds a Bachelor of Science degree in Business, majoring in Business Information Technology with a minor in International Business from the Pamplin College of Business, Virginia Tech. Matt is employed as an Analytics Consultant in Deloitte's Washington D.C. Federal practice.

Alan S. Abrahams is an Associate Professor in the Department of Business Information Technology, Pamplin College of Business, Virginia Tech. He received a PhD in Computer Science from the University of Cambridge, and holds a Bachelor of Business Science degree from the University of Cape Town. Dr. Abrahams' primary research interest is text mining for defect discovery. He is a Senior Editor of Decision Support Systems, and has published in a variety of journals including *Production and Operations Management*, *Decision Support Systems*, *Expert Systems with Applications*, *Journal of Computer Information Systems*, *Communications of the AIS*, and *Group Decision and Negotiation*.

Rich Gruss is a PhD candidate in Business Information Technology, at the Pamplin College of Business, Virginia Tech, and Lead Programmer of Virginia Tech's Math Emporium. Rich holds an MBA degree from Loyola University of Chicago, and an MS degree in Information and Computer Science from UNC Chapel-Hill. Rich's primary research interest is in text analytics.

Johnathon P. Ehsani is an Assistant Professor in health and policy and management at Johns Hopkins Bloomberg School of Public Health. He was previously an injury prevention researcher at the Eunice Kennedy Shriver National Institute of Child Health and Human Development. He received his M.P.H. from the University of Sydney, and his Ph.D. from the University of Michigan. His research focuses on preventing injuries, particularly during childhood and adolescence. He has conducted research in Australia, China, India, and the United States.

Please cite this article as: M. Winkler, et al., Toy safety surveillance from online reviews, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.016