Safety Science 104 (2018) 260–268

Contents lists available at ScienceDirect

Safety Science

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

Social media analytics for quality surveillance and safety hazard detection in baby cribs

Vaibhav Mummalaneni[a,*], Richard Gruss[a], David M. Goldberg[a], Johnathon P. Ehsani[b], Alan S. Abrahams[a]

[a] Department of Business Information Technology, Pamplin Hall, Suite 1007, Virginia Tech, 880 West Campus Drive, Blacksburg, VA 24061, United States
[b] Johns Hopkins Bloomberg School of Public Health, Department of Health Policy and Management, Center for Injury Research and Policy, 624 N. Broadway, Suite 555, Hampton House, Baltimore, MD 21205, United States

ARTICLE INFO

Keywords:
Baby cribs
Defect discovery
Online reviews
Text mining
Sentiment analysis

ABSTRACT

Defects in baby cribs and related products can cause injuries and deaths, and they cost manufacturers and distributors millions of dollars in fines and legal fees and even more in losses of sales and brand image. There has been no prior research regarding automated defect discovery from online reviews of baby cribs, and prior safety defect discovery methods designed and calibrated for other industries must be adapted. We aim to determine which words and phrases are indicators of defects in online reviews and whether sentiment analysis is sufficient for automated defect discovery in the baby crib industry. We find that sentiment analysis serves as a useful tool for automated defect discovery in the baby crib industry and create a supplementary set of "smoke terms" that are strong indicators of safety defects in online reviews of baby cribs. Using our term-based scoring method, we observe a 59% improvement in precision and a 60% improvement in recall when compared to the top-performing prior sentiment method. Our findings provide actionable insights into how analysis of online reviews and other social media can improve baby crib quality management techniques. These terms can be used with immediate effect to monitor and more rapidly identify defects and rectify them before injuries or deaths occur.

1. Introduction

Baby crib safety and defects are a major concern for families, crib manufacturers and distributers. Regulators and consumers are very sensitive to potential safety risks and therefore even just a few cases of defects or safety hazards can lead to a mass recall, costing companies millions of dollars and causing consumers losses of time, money, and peace of mind. Between 1990 and 2008, an estimated 181,654 children younger than two years of age were treated in emergency departments in the United States for injuries related to cribs, playpens, and bassinets, 83.2% of which involved cribs (Yeh, 2011). Between January and September of 2015, there were 6 separate recalls of baby cribs and crib mattresses cited by the Consumer Product Safety Commission (CPSC), involving more than 300,000 individual units (CPSC, 2015). In 2009, the CPSC recalled 2.1 million "drop-down" cribs made by Stork Craft after it was found that they could trap and sometimes injure or kill children. These cribs were sold by many major retailers, such as Amazon, Target, Sears, and Wal-Mart (Smith and Rooney, 2009). In addition to suffering from a tarnished reputation, the manufacturer was required to provide consumers with free repair kits for the cribs, and pay for legal fees.

The industry currently attempts to self-regulate to avoid mass recalls such as the Stork Craft example. After the 2009 incident, companies such as Toys R' Us phased out drop-down cribs, while other companies started manufacturing different types or tried to improve their designs. Companies also frequently recall their own products due to safety concerns and often offer free repairs or replacements for faulty products.

Prior work in hazard discovery from online reviews has studied other industries, such as motor vehicles (Abrahams et al., 2015; Abrahams et al., 2012) and children's toys (Winkler et al., 2016). Academic research regarding crib defect discovery has been minimal, though companies have likely done their own industrial research into the subject. This paper will attempt to provide a more proactive method for finding baby crib safety concerns by analyzing online consumer reviews. Using this methodology, firms can potentially find possible defects or safety concerns while they are relatively small in magnitude and avoid huge scandals and recalls.

This paper is structured as follows. First, we motivate the need for crib quality management research targeted specifically at defect

* Corresponding author.
E-mail address: vaibhav1@vt.edu (V. Mummalaneni).

https://doi.org/10.1016/j.ssci.2018.01.003
Received 10 January 2017; Received in revised form 29 October 2017; Accepted 8 January 2018

discovery from online reviews. Next, we discuss related work. We describe our contributions and the research questions we aim to address. We lay out our process for quality surveillance in the crib industry through consumer review analysis. We discuss and evaluate the application of our crib defect discovery and classification approach using a large data set of Amazon reviews. Finally, we summarize our conclusions and propose further research.

2. Background and related work

In this section, we discuss related work on sentiment analysis, electronic word of mouth, text mining, and quality management. We then review the coverage and limitations of this prior work.

2.1. Sentiment analysis

Sentiment analysis is a process that allows users to mine text and find out whether the emotion of the content is positive or negative by, for instance, comparing each word to a lexicon of positive and negative words. Abbasi et al. examined online forum messages to determine their sentiment (positive or negative) (Abbasi et al., 2008). This method can be used to determine which features of a product cause dissatisfaction in customers, which can then be used to identify product defects. Other studies have applied this methodology to financial markets to predict volatility (Antweiler and Frank, 2004).

Some in the sentiment analysis field presume that postings with a high negative valence indicate a defect in the product. However, this is not always the case, as sometimes a posting that is largely negative may merely report a nuisance or poor product design rather than an actual defect in the product. Particularly in the case of baby cribs, parents of young children are often very stressed, and this could lead to a great deal of sentiment polarity in their comments, even if the problem with the product is not a large or generalizable issue outside of their particular context.

Due to these drawbacks, generic sentiment polarity analysis may not yield proper results when applied to baby cribs. A reviewer might post a seemingly more negative comment about a bad assembly manual than a defect with a hinge if it did not directly affect their child. Therefore, defects must be prioritized based on their potential safety threat: a major defect, such as a crib that is prone to collapsing, must be differentiated from a minor defect, such as a squeaky crib door.

2.2. Electronic word of mouth

There has been a great deal of research regarding electronic word of mouth and reviews, mostly regarding their effect on sales and marketing. Amblee and Bui investigated the effect of electronic word of mouth (eWOM) in the e-book industry (Amblee and Bui, 2011). Applied specifically to Amazon Shorts e-books, they found that eWOM can affect the reputation of the product (the book), the brand (the author), and the reputation of complementary goods (books in the same or similar category).

Other studies have examined eWOM fragments to identify some linguistic patterns. One study found that the text in online reviews of cameras tends to convey strong emotional arousal, implying that usually only people with very strong feelings, positive or negative, write online reviews for products (Pollach, 2006). The same study also found that online reviews tend to be written in a much more professional and serious manner than most other online communications. These two points are important when applying eWOM analysis to defect discovery because it could imply that small defects that do not arouse strong emotions may not be written about, even if they are latent and potentially dangerous. It also could imply that the defects that are written about could be significant, but because reviewers may attempt to write like unbiased critics, the wording may be less polarizing than in other online written content.

2.3. Text mining

Text mining has frequently been applied to e-mails, news articles, online forums, and customer reviews to garner business intelligence, often to support business' decision making. For instance, text mining has been used to classify inbound emails as complaints or non-complaints (Coussement and Van den Poel, 2008). It has also been applied to unstructured consumer generated content to identify consumer issues (Spangler and Kreulen, 2008). Other researchers have used text mining to analyze consumer reviews and forecast box-office success for films (Duan et al., 2008).

Although there have been many studies regarding using text mining to gain competitive business intelligence, these techniques have only recently been applied to product quality and defect discovery (Pan et al., 2014; Vallmuur, 2015). Further exploration is required to extend prior works to new applications and to maximize the performance of the data mining techniques. Abrahams et al. have applied text mining and analytics to defect discovery in the automotive industry (Abrahams et al., 2012) and then subsequently created a framework for defect discovery using text mining across industries (Abrahams et al., 2015). This framework has been adapted to uncover safety hazards in children's toys (Winkler et al., 2016), as well as performance defects in dishwasher appliances (Law et al., 2017). In this study, we attempt to adapt and improve these methods, with specific focus on safety defect discovery in the baby crib industry.

2.4. Quality assurance in baby crib manufacturing

Quality assurance in the baby crib manufacturing industry includes techniques focused on both the supply-side and the demand-side. On the supply-side, firms use traditional product testing, and Statistical Process Control (SPC). Product testing consists of stress-testing the product to identify how and why products fail. This aids in future product improvement as well as identifying common areas of failure in the current product. Statistical Process Control (SPC) utilizes statistical analysis to monitor and control quality during the manufacturing process. It primarily relies upon "acceptance sampling, statistical process monitoring and control, design of experiments, and capability analysis" (Woodall and Montgomery, 1999). Acceptance sampling is used to make decisions regarding "lots" (production batches). Firms use statistical process monitoring to detect changes or abnormalities in the manufacturing process. They can also design experiments to identify which factors of the manufacturing process have the largest effect on product quality. Finally, firms can use capability analysis to determine whether a process is capable of meeting producer or consumer quality requirements (Woodall and Montgomery, 1999).

On the demand-side, firms monitor consumer feedback via surveys (such as Consumer Reports consumer surveys) or direct consumer responses (such as complaint call center hotlines or e-mail complaints). These methods are utilized after the products have been produced and are more reactive than the supply-side quality assurance techniques. Using demand-side and supply-side techniques in tandem creates a more holistic analysis for the firm.

2.5. Summary

Most prior work in the fields of sentiment analysis, electronic word of mouth analysis, text mining, and social media surveillance has not dealt with their applications specifically to defect discovery. For the few studies that have undertaken defect discovery (Abrahams et al., 2015; Abrahams et al., 2012; Law et al., 2017; Winkler et al., 2016), none have assessed the baby crib industry. We aim to adapt, apply, and verify these techniques of analyzing social media and online reviews for the baby crib industry, where we believe these methods could provide valuable feedback to manufacturers and regulators and also yield safer outcomes for parents and their children.

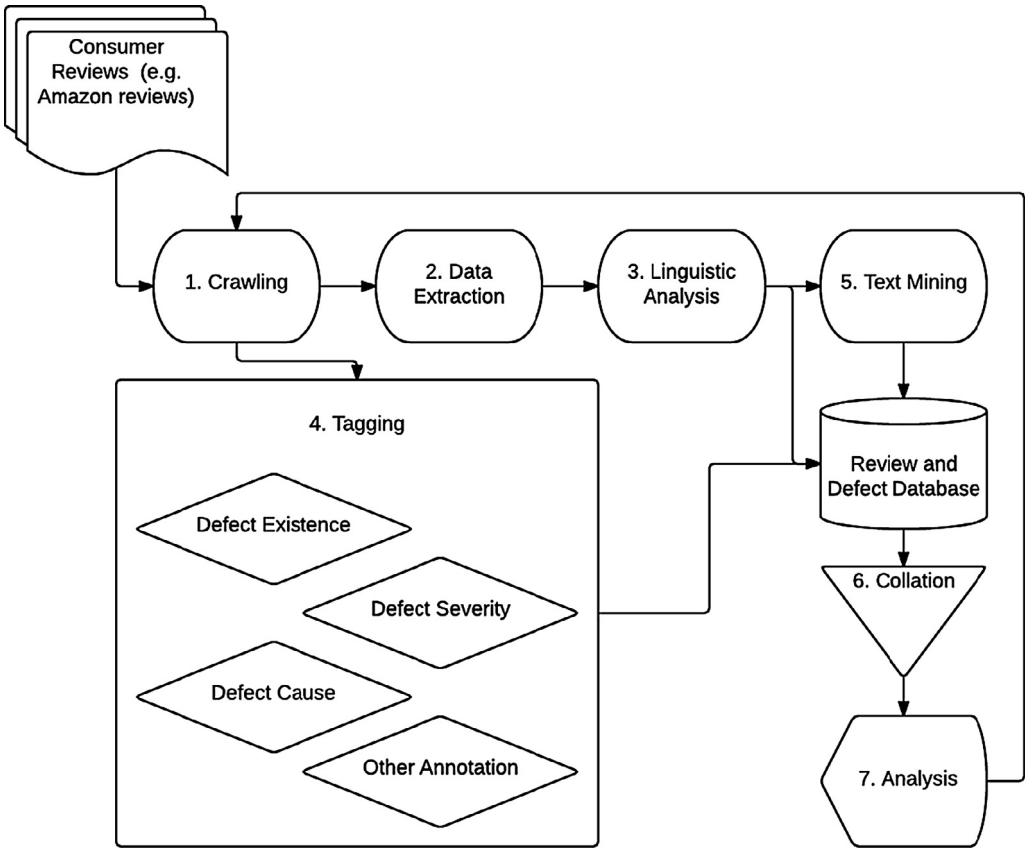

Fig. 1. Process for crib quality management using online consumer reviews.

3. Research questions and contributions

In this paper, we attempt to answer a few research questions. First, how widespread are safety defects in the baby crib and related products industry, and how can these defects be categorized? In detecting these safety defects, we seek to define a series of "smoke terms," or words and phrases that appear especially prevalently in online discussions of safety defects as opposed to other discussions. As such, we ask a second research question: what smoke terms in online reviews are indicators of defects in baby cribs? Third, is sentiment analysis using online reviews sufficient for automated detection of defects in the crib industry? Finally, we ask if there is variation in the rate of defects identified according to the product brand?

4. A text mining framework for baby crib industry quality surveillance

Fig. 1 shows a process model for crib quality management from social media. The process begins with crawling (1.) of Amazon Reviews to gather customer feedback; this process would also work using reviews obtained from other sources such as Babies R' Us, Wal-Mart, Target, or other crib retailers. Next, we extract data (2.) from these posts, including username, date of review posting, and the text of the review itself. Third, we perform linguistic analysis (3.) to analyze the text of the reviews for content and sentiment. We apply existing sentiment methods such as AFINN (Nielsen, 2011), ANEW (Bradley and Lang, 1999), and Harvard General Inquirer (Kelly and Stone, 1975) in order to build an initial understanding of consumer feedback. AFINN is a dictionary of 2,477 positive and negative words, identified from Twitter postings, and scored by a single rater. Each word is associated with a sentiment strength varying from +1 to +5 (for positive words) and varying from −1 to −5 (for negative words). The distribution of word sentiments in AFINN is bimodal, with modes at −2 and +2. In ANEW, a list of English words were individually presented to psychology students and rated on a 9-point scale for pleasure or displeasure, with higher scores (close to 9) indicating greater pleasure, and lower scores (close to 1) indicating greater displeasure. Harvard General Inquirer segments English words into dozens of different semantic categories, with category membership being binary (0 or 1), rather than continuous. Amongst the General Inquirer word categories are sentiment-related categories, such as words "of Positive outlook" (1,915 words) and words "of Negative outlook" (2,291). We employ each of AFINN, ANEW, and Harvard General Inquirer to score the reviews in our dataset for sentiment. Next, we tag (4.) a sample of the reviews with whether or not they identify a defect, how severe the defect is, what the cause of the defect is, and any other pertinent information. "Smoke terms" are words or phrases that are unusually prevalent in reviews pertaining to safety defects (Abrahams et al., 2012). We create a domain-specific list of smoke terms by identifying words and phrases that are substantially more common in baby crib reviews that mention safety defects than in baby crib reviews that do not mention safety defects. We apply the Correlation Coefficient (CC) information retrieval algorithm (Fan et al., 2005), which ranks words or phrases based on their prevalence in safety defect reviews as opposed to non-safety defect reviews, to obtain smoke term lists. Using the top words or phrases identified by the CC scoring algorithm weighted by each term's CC score, we then employ the smoke term list to score the remaining untagged reviews (5.), and obtain an assessment of the likelihood that each review discusses a defect. The information from the data extraction (2.), the linguistic analysis (3.), the tagging (4.) and the text mining (5.) are stored in a review and defect database. Then, the threads are collated (6.) to allow for analysis (7.).

5. Methodology

For the construction of smoke term lists we utilized Amazon.com

consumer reviews: we identified 12,742 baby crib reviews from Amazon.com's set of all reviews from the years May 1996 through July 2014, filtering to include only those reviews under the category of "Baby Products, Nursery" (McAuley et al., 2015). Using this source, we created a number of alternative smoke term lists, as candidates for assessment:

1. ***Unigram*** *(single word) smoke lists created from manually tagged online reviews:* We used the full selection of Amazon reviews of baby cribs and related products. A group of undergraduate business students tagged these reviews to classify their defect status, using the tagging protocol shown in Appendix A. Reviews were randomly selected for display, and each tagger was presented 500 reviews. A total of 12,742 reviews (the full data set) were tagged in this manner. We partitioned the reviews using a 50–50 split and randomly selected half of the 6,371 reviews as our training set. Using these tagged reviews from the training set, we analyzed which unigrams (single words) were most associated with defects, using a correlation coefficient (CC) analysis (Fan et al., 2005). The lead researcher then reviewed the most prevalent words ranked by the CC analysis, and created a list of the 200 most appropriate unigrams to use in the dictionary. This list constituted the Baby Cribs Unigram smoke list.
2. ***Bigram*** *(2-word phrase) smoke list created from manually tagged online reviews:* The process in the previous item was used, except using bigrams (two word clusters) to create the Baby Cribs Bigram smoke list.
3. ***Trigram*** *(3-word phrase) smoke list created from manually tagged online reviews:* The process in the previous item was used, except using trigrams (three word clusters) to create the Baby Cribs Trigram smoke list.

5.1. Data coding

For classifying defects in reviews, we used two domain-independent classification schemes, defect severity and injury timing. The coding protocol, along with samples of each code, are shown in Appendix A.

Note that the term "safety defect" is used for convenience throughout this paper to refer safety concerns expressed by consumers that have been noted by the tagging team, and should *not* be taken to imply that a defect has been confirmed by the manufacturer or any other party.

5.2. Training set

Due to the large volume of reviews and the intermittent availability of tagging team members, we scored the full set of 12,742 reviews across three tagging sessions. The first session tagged 3,701 random reviews, the second session contained 1,539 reviews, and the final session tagged the remaining 9,224 reviews. Some reviews were tagged multiple times by design to make sure scores were reliable and that inter-rater reliability was robust. Cohen's κ (Cohen, 1960) was > 0.75 for Defect Severity (N = 1,592 cases; 1,396 agreements; 196 disagreements; 88% percentage agreement), indicating substantial inter-rater reliability on that attribute. As Cohen's κ was unsatisfactory for Injury Timing, we excluded Injury Timing from further analysis. To reconcile the tags of reviews with multiple taggers, we used a majority conservative vote system. We took the majority vote of the taggers, and if votes were tied, then we took the most conservative decision (safety defect over performance defect over no defect). We then took the reviews and partitioned them using a 50–50 split to create the training set and the holdout set, each containing 6,371 reviews.

5.3. Holdout set

The holdout set contained all the reviews that were not included in the training set. The holdout set contained 6371 reviews. Following are three sample reviews containing potential safety defects, identified in the holdout set:

> *"Although this bed is beautiful it is a NIGHTMARE! The bed was simple to put together but we had trouble getting some of the screws / bolts to match up. Total pain! My biggest complaint are the BOWS!! BEWARE!! They are so sharp that my 2½ yr old keeps waking up crying from bumping her head on them!!! Also, the side rail are not tall enough to keep her from falling out … which she occasionally does. I am not a fan of this bed!!"*

> *"I bought this crib for my first child, brought it home set it up and within a week of her being in the crib. The bottom bits of the drop part of the railing were coming out and the railing broke. It ended up being unsafe for her and I had to buy another one with better support."*

> *"Long story short, we needed a new crib mattress. This doesn't cut it. I gave it a few days to fill out after having been vacuum - packed, but it's too squishy for an infant, let alone my baby who rolls to his belly. He would suffocate – there's that much give to the mattress. Comfy for adult tastes, yes, and even big kids, but not safe for infants in a crib."*

5.4. Validation set

We used the holdout set of 6,371 "Baby Products, Nursery" reviews to test multiple review scoring methods. The holdout set was scored using three *sentiment* methods (ANEW (Bradley and Lang, 1999), AFINN sentiment words (Nielsen, 2011), and the Harvard General Inquirer Negative word list (Kelly and Stone, 1975) and was also scored using the three *smoke term lists* (Unigram, Bigram, and Trigram). We took the top 200 and bottom 200 scoring reviews from each method to create a validation set containing 2400 reviews (top 200, and bottom 200, from each of the 6 methods). The "top 200" reviews represented those that were most likely to indicate a safety defect. Because AFINN has increasingly negative scores for more negative sentiment, we reversed the order of the scores so that the "top 200" AFINN reviews would be those that had the lowest scores and were most likely to indicate a defect. Similarly, ANEW has increasingly positive scores for positive emotion, so we had to reverse the scores so that the "top 200" scores would contain the reviews with the lowest scores indicating the least positive sentiment which would be more likely to be indicative of a safety defect. This process ensured consistency and clarity across the different scoring methods so that "top" would always indicate a high likelihood of a safety defect and "bottom" would always indicate a low likelihood of a safety defect. A visual representation of the review partitioning is shown in Fig. 2: the numbers 1, 2, 3, 4 in black shaded circles in Fig. 2 indicate the sequence of four steps described above: (1) creating the training set, (2) creating the holdout set, (3) scoring the reviews using each of the six scoring methods, and (4) assembling a validation set by finding the top and bottom 200 reviews for each of the six scoring methods.

6. Results

In this section, we describe the baby crib-specific smoke terms identified by our analysis, and evaluate the performance of sentiment analysis and smoke term scoring on our data set.

6.1. Baby crib-specific smoke term lists

For illustration, Table 1 reports the twenty unigrams, bigrams, and trigrams that were most prevalent in safety defects in the training set. Table 2 indicates the terms, from these smoke lists, that were most prevalent in the holdout set.

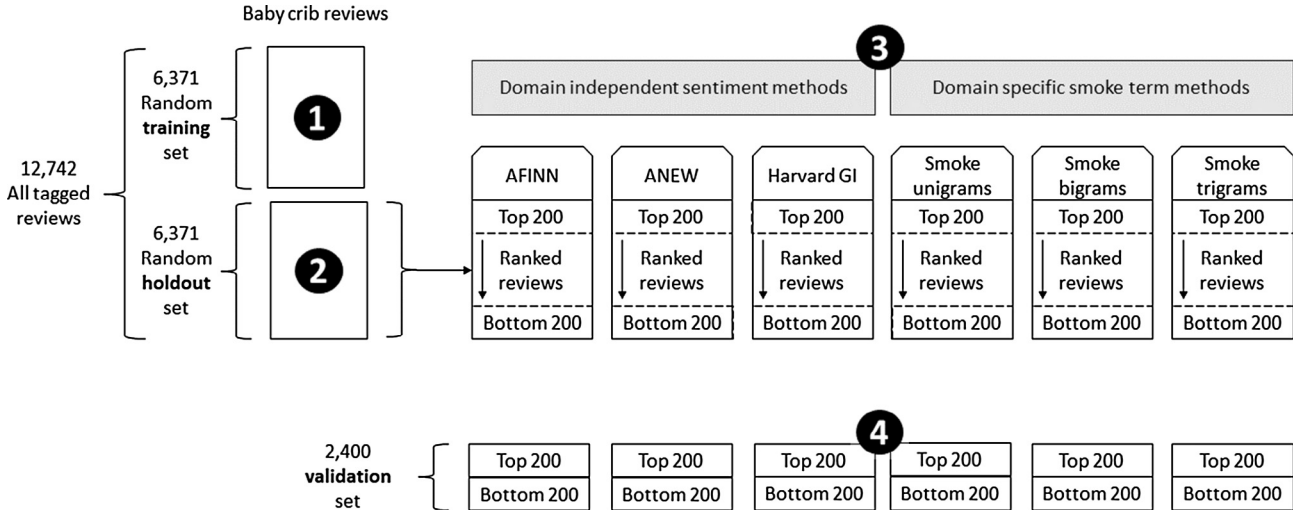

Fig. 2. Process for data set partitioning and scoring.

Table 1
Unigrams, bigrams, and trigrams most prevalent in the *training* set.

Unigram	Bigram	Trigram
dangerous	choking hazard	work for my
choking	one drawback	the spinning balls
sharp	return policy	the ground when
spinning	the spinning	these signs are
munched	spinning balls	put anything heavy
hazardous	dangerous for	is very active
recalled	any weight	stable enough for
lungs	reviewers were	a very slippery
disconcerting	base it	it once it
decently	materials is	problem we have
stabilizer	we threw	not stable enough
policy	very slippery	chair is too
threw	caps are	this chair should
corresponding	past i	as mine did
broke	nearly all	no longer have
unsafe	please spend	would not snap
hazards	mine did	the rails i
harm	square its	a better solution
stripped	so by	not move and
exposing	time ill	the other 3

Table 2
Smoke terms most prevalent in the *holdout* set.

Unigram	Bigram	Trigram
not	of the	one of the
out	choking hazard	a choking hazard
off	a choking	of the panels
choking	the manufacturer	out in the
child	return policy	in his mouth
broke	dangerous for	the wrong place
apart	we heard	the bottom of
hazard	the panels	
dangerous	his mouth	
return	not snap	
sharp	a flat	
mouth	wrong place	
safety	his head	
recalled	safety is	
threw		
unsafe		
hazards		
poorly		
potential		
trash		

Table 3
Count of reviews containing safety defects, per scoring method.

	Dictionary	NO safety defect	Safety defect	Grand total
SENTIMENT	**AFINN**			
	Bottom 200	197	3	200
	Top 200	180	**20**	200
	ANEW			
	Bottom 200	186	14	200
	Top 200	198	**2**	200
	Harvard GI Negative			
	Bottom 200	198	2	200
	Top 200	183	**17**	200
SMOKE	**Unigram**			
	Bottom 200	196	4	200
	Top 200	165	**35**	200
	Bigram			
	Bottom 200	199	1	200
	Top 200	165	**19**	200
	Trigram			
	Bottom 200	198	2	200
	Top 200	165	**15**	200
	GRAND TOTAL	2230	**170**	2400
	Baseline (expected per 200 random reviews)	195	**5**	200

6.2. Sentiment analysis and smoke term scoring for defect discovery

We analyzed the top and bottom 200 reviews that were scored by each scoring method (2,400 reviews in the validation set). The count of safety concerns identified by each method is shown in Table 3. The number of safety defects in the top 200 scored reviews for each method is shown in **bold**, to draw attention to how well the method is able to identify potential safety hazards in its top 200-ranked reviews.

There were a total of 158 safety defects in the 6,371 reviews in the holdout sample (2.48%) - which corresponds to about 5 defects per 200 reviews. 5 out of 200 can therefore be regarded as the "baseline" performance threshold: any successful method should perform better than this baseline, since any successful method should find significantly more defects in its top-scoring N reviews than the expected number of safety defects in a random selection of the same number of reviews. The baseline (for $N = 200$) is indicated in the final row of Table 3, for comparison.

Next, we needed to verify that there were statistically significantly more defects in the top 200 reviews versus the bottom 200 reviews. We

Table 4
Comparison of proportion of safety defects found in top versus bottom scoring reviews, for each scoring method.

Scoring method	p-value (Chi-test)	Interpretation
AFINN	< .001**	Top 200 have statistically more defects than Bottom 200
ANEW	< .001**	Bottom 200 have statistically more defects than Top 200
Harvard GI Negative	< .001**	Top 200 have statistically more defects than Bottom 200
Unigram	< .001**	Top 200 have statistically more defects than Bottom 200
Bigram	< .001**	Top 200 have statistically more defects than Bottom 200
Trigram	< .001**	Top 200 have statistically more defects than Bottom 200

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

used a Chi-squared test to compare the number of defects found in the bottom and top of each dictionaries scored reviews. The results are shown in Table 4.

All findings were in the direction expected, with the exception of ANEW, where high scoring reviews (bottom 200) with the most positive sentiment were associated with safety defects. An analysis of the reason for this revealed that “bed”, “toddler”, and “baby” – highly positive words in the ANEW word rankings – were mentioned frequently in safety defects, and contributed to the highly positive purported sentiment of these reviews according to ANEW.

Given these findings, we then tested to see which of these scoring methods were better at finding defects than randomly selecting reviews. We then used a Chi-squared test to determine which scoring methods outperformed the baseline (of 5 safety defects per 200 reviews) for their top 200 scoring reviews. The results are shown in Table 5.

Next, we determined whether the safety defects versus non-defects for each approach have different scores for each scoring method. We ran a *T*-test to check whether each method significantly higher or lower scores on safety defects or non-defects. The results are shown in Table 6.

As we delineate between safety defect reviews and non-safety defect reviews based on several scoring systems, it is useful to compute common machine learning metrics such as precision, recall, and lift. We define precision as the proportion of safety defect predictions made that are correct; recall as the proportion of safety defects detected out of all available safety defects in the holdout set; and lift as the ratio of observed precision to the level of precision expected by random chance. Unlike binary classification methods that assign a predicted class to each review, we instead rank each review by each method’s estimate of its likelihood of reflecting a safety defect. Therefore, no unique precision, recall, or lift scores totally encompass the performance of any of our scoring methods. Instead, we compute precision, recall, and lift metrics by choosing several cutoffs for the top *N*-ranked reviews. For example, taking the top 200-ranked reviews according to our unigram smoke term scores, the precision is 35/200 as 35 of the top 200 reviews reflect safety defects. This constitutes 35/158 total safety defect reviews in the holdout sample, or recall of 0.222. Finally, the lift for the top 200 reviews for the unigram smoke scoring method is computed as the precision of 0.175 divided by the baseline rate of 0.0248 – that is, lift = 7.056 = 0.175/0.0248 for the top 200 reviews for the unigram smoke scoring method. A simple interpretation of the lift figure of 7.056 is that the Smoke Unigram method finds 7 times more safety defects in its top 200 reviews than what we would expect to find in a random sample of 200 reviews. We present the precision, recall, and lift scores for our six methods at five cutoffs in Table 7. Additionally, we present the baseline threshold, or the levels of precision, recall, and lift that would be expected by random chance. Finally, for ease of visual interpretation, we bold the best precision, recall, and lift scores at each cutoff.

As Table 7 demonstrates, the smoke term unigrams, bigrams, and trigrams outperformed all of the pre-existing scoring methods. At lower cutoffs, the smoke term unigrams outperformed all of the other methods, but at higher cutoffs, the smoke term bigrams and trigrams actually surpassed the smoke term unigrams. At lower cutoffs, precision scores are higher, but recall scores are lower; and at higher cutoffs, precision scores are lower, but recall scores are higher. The choice of which cutoff to use rests with the user, but when choosing a middle-ground such as classifying the top 200-ranked reviews as safety defects, we see that the unigram smoke term method offers a 59% improvement in precision and a 60% improvement in recall compared to AFINN, the top-performing traditional alternative. These findings indicate strong evidence that industry-specific smoke terms are better predictors of safety defects in online reviews of baby cribs than any of the pre-existing sentiment-based scoring methods.

Finally, we created a “lift chart” to display how the smoke term scoring methods compared to the sentiment methods across the full range of possible cutoffs. For the construction of the lift charts, all reviews in the holdout set are ranked by each scoring method, from highest ranked to lowest ranked (X-Axis). The Y-Axis then shows how many safety defects were found in the top *N*-ranked reviews. The lift chart shows that the unigram smoke term method is the best standalone scoring method (the curve bulges highest above the diagonal line, meaning that it finds the most defects the fastest in its top ranked reviews). The lift chart is shown in Fig. 3 – for legibility, only three methods are shown (the unigram smoke term method, which generally performed better than bigrams and trigrams; plus the top two sentiment methods: AFINN and Harvard GI Negative).

6.3. Brand-level analysis

As a final step in our analysis, we sought to examine the extent to which there are substantial differences in safety defect smoke term scores between brands in the baby cribs industry. This way, we may determine if all brands offer relatively homogenous levels of safety or if

Table 5
Comparison of proportion of safety defects found by each scoring method, to baseline proportion of defects in holdout set.

Scoring method	p-value (Chi-test)	Interpretation
AFINN	< .001**	Top 200 have statistically more defects than baseline
ANEW	.17	Top 200 *not* statistically more defects than baseline
Harvard GI Negative	< .001**	Top 200 have statistically more defects than baseline
Unigram	< .001**	Top 200 have statistically more defects than baseline
Bigram	< .001**	Top 200 have statistically more defects than baseline
Trigram	< .001**	Top 200 have statistically more defects than baseline

** Indicates statistical significance at the 99% confidence level

Table 6
Statistical difference in means, for defects versus non-defects, for each scoring method.

Scoring method	Mean score		p-value (*T*-test)	Interpretation
	Not safety defect	Safety defect		
AFINN	11	−1	< .001**	Safety defects have a significantly lower score
ANEW	117	225	.177	Safety defects do *not* have significantly different score
GI Negative	6	11	< .001**	Safety defects have a significantly higher score
Unigram	21,254	46,062	< .001**	Safety defects have a significantly higher score
Bigram	21,079	49,900	< .001**	Safety defects have a significantly higher score
Trigram	21,137	48,552	< .001**	Safety defects have a significantly higher score

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

Table 7
Precision/recall/lift scores of top *N*-ranked reviews for each scoring method.

Cutoff	Baseline	AFINN	ANEW	Harvard GI	Smoke unigrams	Smoke bigrams	Smoke trigrams
100	0.025/0.016/1.000	0.120/0.076/4.839	0.000/0.000/0.000	0.060/0.038/2.419	**0.210/0.133/8.468**	0.120/0.076/4.839	0.080/0.051/3.226
200	0.025/0.031/1.000	0.110/0.139/4.436	0.01/0.013/0.403	0.075/0.095/3.024	**0.175/0.222/7.056**	0.095/0.120/3.831	0.075/0.095/3.024
500	0.025/0.078/1.000	0.086/0.272/3.468	0.008/0.025/0.323	0.074/0.234//2.984	**0.122/0.386/4.919**	0.096/0.304/3.871	0.076/0.241/3.065
1,000	0.025/0.157/1.000	0.061/0.386/2.460	0.008/0.051/0.323	0.056/0.354/2.258	**0.096/0.608/3.871**	0.071/0.449/2.863	0.076/0.481/3.065
2,000	0.025/0.314/1.000	0.043/0.544/1.734	0.013/0.158/0.504	0.048/0.601/1.915	0.060/0.759/2.419	**0.061/0.772/2.460**	**0.061/0.772/2.460**

some brands are potentially much safer than others. Within our holdout sample of 6,371 online reviews, we detected all unique brand names, and we averaged the smoke term scores for each brand. We removed any brands with under 50 online reviews to ensure that the averages were based on sufficiently large and representative samples. Table 8 displays statistics for the 18 remaining brands.

Interestingly, our analysis reveals enormous differences between brands' unigram smoke term scores. Brand 11 had by far the highest average smoke term score, with a score 66% higher than that of the next highest brand, Brand 4, and 359% higher than that of the lowest-scoring brand, Brand 10. The average unigram smoke term scores were largely consistent with the rate of safety defects observed. We computed a 0.754 correlation between the average unigram smoke score for each brand and the safety defect rate for each brand, indicating substantial agreement between our unigram smoke score method and the presence of safety defects in each brand. Therefore, it seems evident that there is considerable brand-level variability in the presence of safety defects in the baby cribs industry.

7. Discussion

In this paper, we assessed safety defect discovery for baby cribs using a selection of six scoring methods. We used a limited number of reviews in our tagging process due to physical and time limitations. In the future, the variety of scoring methods and the number of tagged reviews could be expanded. Also, human taggers could have incorrectly tagged some of the reviews based on incorrect interpretations. The methodology we used is inherently susceptible to these kinds of mistakes, so additional investigations could be undertaken to supplement our analysis. As an extension to our analysis, future work could also apply our methodology to other forms of social media such as Twitter or Facebook posts as well as posts in relevant blogs and forums.

In the future, we would like to do variants of the analyses using other manually created dictionaries such as term lists created from the US CPSC NEISS database of hospitalizations, US CPSC Product Recall reports, and the European Commission Rapid Alert System (EC RAPEX) data set.

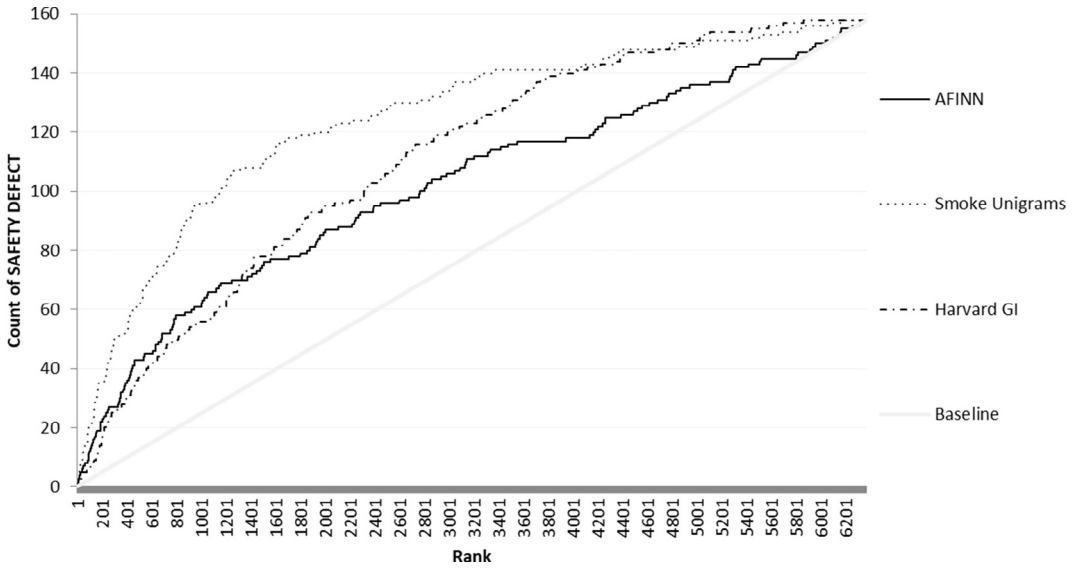

Fig. 3. Lift chart: number of safety defects found in top-ranked items for each scoring method.

Table 8
Average unigram smoke term scores for most-reviewed baby crib brands.

Brand	Review count	Average unigram smoke term score	Tagged safety defects	Safety defect rate
Brand 11	170	16,164	16	9.41%
Brand 4	54	9,725	5	9.26%
Brand 15	168	9,712	14	8.33%
Brand 5	116	9,290	10	8.62%
Brand 13	61	8,938	1	1.64%
Brand 14	56	8,854	3	5.36%
Brand 17	52	8,377	2	3.85%
Brand 9	97	8,263	1	1.03%
Brand 3	201	7,913	2	1.00%
Brand 6	239	7,247	7	2.93%
Brand 8	142	7,162	3	2.11%
Brand 16	66	6,998	0	0.00%
Brand 18	91	6,900	3	3.30%
Brand 12	693	6,767	18	2.60%
Brand 7	240	6,753	8	3.33%
Brand 1	104	4,867	1	0.96%
Brand 2	394	3,738	0	0.00%
Brand 10	319	3,517	2	0.63%

8. Implications for practice and research

The findings of this study have numerous implications for injury prevention practitioners:

- Defects are mentioned relatively frequently in Amazon reviews of baby cribs and related products, but the majority of reviews do no mention any safety defect. This implies that injury prevention practitioners would benefit from using an automated Defect Discovery System such as ours to mine reviews to find relevant ones and take preventive action.
- The smoke terms that we identified in this study strongly predict the existence of safety defects in baby cribs and related products. This implies that practitioners would benefit from using web crawlers in addition to our analysis tools to scan Amazon reviews for defects in their baby cribs as a safety surveillance tool. This could be applied to other forums or social media sites in the future.

For industry, the implications are as follows:

- Smoke term analysis is also useful as a competitive tool. Brands may assess their product safety relative to competitors and use this information to identify deficiencies and develop a competitive advantage. Additionally, new market entrants can seek out these factors to identify and mitigate important risk factors.

For researchers, the implications are as follows:

- We have determined that generic sentiment words are relevant to defect discovery in baby cribs. However, more domain-specific smoke term lists, such as those that we identified in our study, offer noticeably better results.
- We have adapted earlier methods for automated defect discovery to the baby crib industry and performed several statistical tests to demonstrate the effectiveness of our modified method. Our results show that this method is applicable to the baby crib industry and can discover safety defects with superior recall, precision, and lift relative to prior methods. This methodology could be adapted even further for more industries to assist more broadly in quality control and product management.

For regulators, we note the following additional implication:

- As it is the mission of national safety regulatory agencies to identify products that may need to be recalled, smoke terms may prove a valuable tool in rapidly monitoring online content for potential safety concerns. Our detailed findings have been actively requested by, and shared with, two national safety regulatory agencies: the United States Consumer Product Safety Commission (US CPSC), and Health Canada, the department of the government of Canada with responsibility for national public health. While the confidentiality of investigations precludes our ability to monitor subsequent action, we have received positive feedback that our procedure and findings are helpful contributions to consumer product safety surveillance efforts.

9. Conclusions

In this paper, we adapted text analytics methods for defect discovery to the baby crib industry. By applying sentiment analysis and smoke term discovery to customer reviews of baby cribs on Amazon, we were able to identify safety defects mentioned within the reviews. We compiled domain-specific smoke term lists containing words and phrases most critical to safety defect discovery in baby cribs. This paper has shown that baby crib quality management can be supplemented by analysis of online reviews, and possibly similar textual sources.

Appendix A. Coding protocol for baby crib defects, with examples

Baby crib defects were coded as follows:

1. Defect severity classifies the level of defect found in the product. There are three sub-categories, “no defect”, “safety defect”, and “performance defect”, which are defined as follows:
   a. “No Defect” includes reviews that do not mention any firsthand experience of manufacturing defects. For example, reviews that are completely complimentary of the product or reviews that mention defects that they have heard about but have not experienced themselves. In this case, the customer mentions a possible defect in another product in this review of a crib that seemingly has no defects:

“This bed is fantastic. Our convertible crib was broken in a move, so it could only be used as a crib (and not convert like we did for our first son). Our second son had a horrible problem of getting his knees and chubby thighs stuck in the slats of the crib so we had to move him to a bed at 15 months. He is a shorty, so I was afraid of him being able to get in and out and falling off, etc. This bed was the perfect choice. It is low to the ground, but incredibly and surprisingly sturdy. It is plain wood so you don't have to deal with some cartoon character so it isn't too overwhelming and instead you can just dress it up with different sheet sets and it looks great. I put this together myself (with the help of my 3 y/o) in about 2 h. It would have been quicker without my little “helper”. I do advise that you use an actual Allen wrench and regular wrench and not the ridiculous cheap ones that come in the little tool kit. That's it! Both of my kids love it and my child has had no issues. Please be advised that this bed only holds up to 50 lb, so I would suggest you not lay down in it with your child. I haven't tried testing the weight limit, but I don't feel like finding out if it would hold me.”

   b. “Safety Defects” are defects that have or could lead to injury or death. For example, a choking hazard, unstable crib legs, or bars that can trap users. The following review describes a severe safety defect that one crib owner experienced:

“I have read some of the reviews and I agree with all the problems that the people are having, but the worst was when my daughter got her legs stuck in the bars. I awoke one morning to her my 1 year old daughter screaming at the top of her lungs and ran into see what was the matter. When I got in her room I seen both legs stuck in between the rails. I ran over to her and struggled to free them. It took me

about 5 min to free them and when I did I took off her PJs to discover both legs were dark purple. She must of stuck like that a little while before she screamed for help. I rubbed her legs till the color came back and seen she had some bruising just around her knees. I was very upset, but I was made to beleive that it was just a freak thing that could never happen again. SHAME ON ME! My daughter is now 17 months, my mother in-law was watching her and while waking from her nap did it again with one leg. In a panic she couldn't free her without dumping lotion on her leg. I am now tring to contact the company. PLEASE beware this can happen."

c. "Performance Defects" are product defects that do not and are unlikely to lead to an injury or death. Examples include misprinted graphics and squeaky hinges.

"i was extremely disappointed in this basket. very flimsy and the lining was cheap and not soft at all. not as pretty as pictured. lining didn't fit right and you could see the velcro where the hood was suppose to attach. i returned it immediately and bought a different basket from another company. would not recommend this to anyone at all."

2. Injury timing classifies the injuries that the writer of the review claims was caused by the product. There are three sub-categories, "no injury, "actual injury occurred", and "potential injury could occur", which are defined as follows:
   a. "No injury" includes reviews that do not mention any injury or any possible injury that could occur due to the product. For example, completely positive reviews and reviews that only mention performance defects with no added safety concern.

"Wish someone would have told me the bumper to the crib set wasn't made for a crib with closed ends…and that the bumper would eventually lose all it's shape and some of the ties would come off in the wash. Cute pattern for baby's room but not a real good value because it doesn't really last. Would have paid more for a better made product and will do so with baby #2."

b. "Actual injury occurred" includes reviews that mention first-hand experience of an injury caused by the product, according to the writer of the review.

" All in all a great product for the price but remember pine is a softer wood so it dents easily and the type of pine is not very dense it feels more like bosa wood… My daughter is now 13 months old and has recently got her little leg and knee wedged in between the slats. It was s bad she had a bruise wrapped around her leg each time."

c. "Potential injury could occur" includes reviews that mention some issue with the product that could cause an injury in the future. For example, durability issues, or broken parts.

"Looks nice and fairly sturdy. […] one drawback that some people may not like is that the bars on the ends are spaced too far apart. Our baby is 9 months old is very active and can pull herself up. couple of times her legs popped through the bars on the end all the way up to her thigh and she was stuck. Not sure why they did not make the spacing smaller like on the sides. the brackets for the mid level point are such that baby tried to use them to try and climb out of the crib. later one when she's stronger they may become stepping points. Only other quirk was 1–2 rough patches on the crib almost cut my fingers while I was wiping it down. had to sand them to make sure baby doesn't scrape herself."

References

Abbasi, A., Chen, H., Salem, H.A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inform. Syst. 26 (3).

Abrahams, A.S., Fan, W., Wang, G.A., Zhang, Z., Jiao, J., 2015. An integrated text analytic framework for product defect discovery. Prod. Operat. Manage. 24 (6), 975–990.

Abrahams, A.S., Jiao, J., Wang, G.A., Fan, W., 2012. Vehicle defect discovery from social media. Decis. Support Syst. 54 (1), 87–97.

Amblee, N., Bui, T., 2011. Harnessing the influence of social proof in online shopping: the effect of electronic word-of-mouth on sales of digital microproducts. Int. J. Electr. Commer. 16 (2), 91–113.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. J. Finance 59 (3), 1259–1294.

Bradley, M.M., Lang, P.J., 1999. Affective norms for English words (ANEW): Stimuli, Instruction Manual and Affective Ratings. Technical Report C-1, Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychosoc. Measure. 20, 37–46.

Consumer Product Safety Commission. 10 Sept. 2015. < www.cpsc.gov > .

Coussement, K., Van den Poel, D., 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. Decision Supp. Syst. 44 (4), 870–882.

Duan, W., Gu, G., Whinston, A.B., 2008. Do online reviews matter? — An empirical investigation of panel data. Decis. Support Syst. 45 (4), 1007–1016.

Fan, W., Gordon, M.D., Pathak, P., 2005. Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. Decis. Supp. Syst. 40 (2), 213–233.

Kelly, E., Stone, P., 1975. Computer Recognition of English Word Senses, North-Holland Linguistic Series.

Law, D., Gruss, R., Abrahams, A.S., 2017. Automated defect discovery for dishwasher appliances from online consumer reviews. Exp. Syst. Appl. 67, 84–94.

McAuley, J., Pandey, R., Leskovec, J., 2015. Inferring Networks of Substitutable and Complementary Products. Knowledge Discovery and Data Mining, Sydney, Australia.

Nielsen, F., 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big Things Come in Small Packages, pp. 93–98.

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. Syst. Res. Behav. Sci. 31 (3), 383–397.

Pollach, I., 2006. Electronic word of mouth: a genre analysis of product reviews on consumer opinion web sites. In: Proceedings of the 39th Hawaii International Conference on System Sciences.

Smith, A., Rooney, B., 2009. Crib Recall: 2.1 Million Deemed Unsafe. CNN Money. 23 Nov. < http://money.cnn.com/2009/11/23/news/companies/crib_recall/ > .

Spangler, S., Kreulen, J., 2008. Mining the Talk: Unlocking the Business Value in Unstructured Information. IBM Press.

Vallmuur, K., 2015. Machine learning approaches to analyzing textual injury surveillance data: a systemic review. Accid. Anal. Prevent. 79, 41–49.

Winkler, M., Abrahams, A.S., Gruss, R., Ehsani, J.P., 2016. Toy safety surveillance from online reviews. Decis. Supp. Syst. 90, 23–32.

Woodall, W.H., Montgomery, D.C., 1999. Research issues and ideas in statistical process control. J. Qual. Technol. 376–386.

Yeh, E., 2011. Injuries associated with cribs, playpens, and bassinets among young children in the US. Pediatrics 479–486.

Further reading

Zhang, Y., Dang, Y., Chen, H., Thurmond, M., Larson, C., 2009. Automatic online news monitoring and classification for syndromic surveillance. Decis. Supp. Syst. 47 (4), 508–517.