Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020

Topic Modeling and Transfer Learning for Automated Surveillance of Injury Reports in Consumer Product Reviews

David M. Goldberg
San Diego State University
dgoldberg@sdsu.edu

Nohel Zaman
Loyola Marymount University
nohel.zaman@lmu.edu

Abstract

Many modern firms and interest groups are tasked with the challenge of monitoring the status and performance of a bevy of distinct products. As online user-generated content has increased in volume, new unstructured data sources are available for mining unique insights. Reports of injuries arising as a result of product usage are particularly concerning. In this paper, we utilize complimentary approaches to address this problem. We analyze two novel datasets; first, a government-maintained dataset of hazard and injury reports and second, a large dataset of cross-industry consumer product reviews manually coded for the presence of hazard and injury reports. We apply an unsupervised topic modeling approach to characterize the hazard and injury reports detected. Then, we implement a supervised transfer learning technique, using information obtained from the government-maintained dataset to detect hazard and injury reports in online reviews. Our results offer improved surveillance for monitoring hazards across multiple industries.

1. Introduction

Product manufacturers employ product prototyping, stress tests, consumer focus groups, and further methods to ensure the quality and safety of consumer products [14]. Yet, according to the National Safety Council, 10.5 million people were treated in emergency departments in 2017 due to consumer product-related injuries. In the modern globalized economy, the breadth of product offerings poses enormous difficulty for firms' surveillance efforts. Large firms can have thousands of distinct models of products to monitor, each sold across the world. In addition to the enormity of this surveillance task, consumer use cases are often difficult to accurately predict prior to a product's sale on the market [25]. Thus, in recent years, many firms have sought to supplement their pre-market product safety efforts with post-market monitoring. For instance, firms may actively monitor warranty claims and product returns to understand patterns underlying consumers' dissatisfaction with product quality.

The urgency of safety surveillance techniques is especially heightened given the enormity of the risk associated with product recalls. If there is sufficient evidence that a product on the market poses substantial risk to consumers or to their property, then federal agencies are obligated to issue a recall of the product. These recalls can be immensely costly for firms, which must reconcile with consumers usually by reimbursing them or offering a free replacement product. In addition, some firms may be subject to millions of dollars of federal penalties [1]. However, the financial impacts of product recalls extend well beyond these initial obstacles. For example, firms affected by prominent recalls can rapidly lose consumers' goodwill as they get a reputation for poor product quality [25]. As such, research has found that firms whose products are recalled experience negative stock returns [21]. Marketing research has found that these firms are in a no-win situation, as efforts to save face by taking a proactive public stance are actually generally counterproductive and reinforce consumers' perceptions that products are unsafe [6].

Hora et al. [14] study the "recall gap", or the difference in time between a product reaching market and its eventual recall. In their analysis of the recall gap over 15 years, the average recall gap ranged from 436 days to 869 days, representing a substantial multi-year period in which hazardous products were sold on the market prior to their eventual recalls. Longer recall gaps are especially dangerous for firms, as subsequent remediation efforts must become more extensive, federal penalties may be higher, and the magnitude of the recall results in a greater loss in goodwill. Thus, any extent to which firms can mitigate against long recall gaps by detecting potential product safety hazards quickly can be enormously beneficial in the long run. Even in less severe situations for which a recall is not necessary, rapid information about the quality and safety of products allows firms to react

URI: https://hdl.handle.net/10125/63866
978-0-9981331-3-3

quickly as they innovate and design future iterations of products.

Given the motivation to rapidly source intelligence on product quality, real-time information is at a premium. For this reason, recent research has utilized online posts, such as social media or online reviews, as a potential source of information to inform these processes. The textual format of these data sources is particularly rich, allowing consumers to post detailed narratives describing their experiences with products. Since the volume of the text data available online is unrealistic for a firm to review in its entirety, automated methods are instead used to efficiently sort and prioritize records for review. Initial research efforts to address this problem have utilized sentiment analysis [16], which rates text on a scale from emotively positive to emotively negative. These studies assume that emotively negative text is most likely to be associated with product safety hazards and search for particularly negative text in hopes of discovering safety hazards. This technique achieved some limited success, but the nuances of language prevent sentiment from capturing the entirety of the problem at hand. For instance, the phrase "my blender blew up" clearly indicates a safety hazard, but none of the words in the phrase are particularly emotively strong.

More recently, studies have sought to use a more nuanced and specialize technique to detect mentions of safety hazards in online content. Researchers curate "smoke terms", or particular words and phrases especially prevalent in online posts that refer to safety hazards [2-4, 12]. These smoke terms may or may not be emotive, differentiating them from sentiment analysis. For instance, the term "airbag" in the automotive industry is non-emotive, but online posts that refer to airbags are very likely to be associated with a safety-related incident in which an automobile crashed [3, 4]. As such, smoke terms have been far more effective as a means for detecting mentions of safety hazards in online media. However, a major limitation of this technique has been that smoke terms are generally limited to a particular industry. While the term "airbag" is an excellent predictor of safety hazards in the automotive industry, it is unlikely that airbags are relevant to many other industries. In the toy industry, for example, the term "airbag" is unlikely to be relevant whatsoever.

This study aims to further the study of safety surveillance using several unique approaches. Our first area of emphasis in this study is that not all safety hazard reports are equal. In some reports, a consumer may state that a product got very hot and could have potentially burned them. While worrying, a firm ought to be much more concerned about a report in which a product caught on fire and burned both the consumer and their property. Thus, we put particular emphasis on the subset of safety hazard reports in which a consumer was injured by a product. Our second area of emphasis is to apply our insights in a cross-industry setting. Rather than limiting our analysis to a single industry of emphasis, our study of product injury reports can span across multiple product categories. To that end, we utilize and label an enormous cross-industry dataset of over 100,000 amazon.com reviews.

We approach this problem using two contemporary text mining approaches that are novel in the safety surveillance literature. First, we apply topic modeling to better understand the distribution of latent topics present in safety hazard reports. We use a large dataset of safety hazard reports maintained by a government agency for this initial stage of text mining. By better understanding not only which topics are likely to be present in safety hazard reports but also which words are likely to be indicative of these topics, manufacturers and interest groups can better understand the nature of product safety in their respective industries and prioritize safety surveillance accordingly. Second, we apply transfer learning to use the large set of safety hazard reports as training data to analyze online reviews. It is difficult to use online reviews as a source of training data for injury reports as injury reports are quite uncommon, thus limiting the availability of a sufficiently large sample. However, using transfer learning, we generate indicative smoke terms using the government data source, and we then reapply these insights in the new domain of online reviews. Doing so allows us to build a high-functioning predictive model whose knowledge is transferred from one domain to another. Using such a model, practitioners can more rapidly sort online posts, prioritizing the most pressing concerns first to mitigate against potential ongoing safety concerns.

2. Literature review

2.1. Product safety data sources

Government agencies such as the Consumer Product Safety Commission (CPSC), Health Canada, the European Union (EU) Health and Safety Authority, and the British Standard Institution (BSI) identify and evaluate risks at different stages of the product safety cycle. These government agencies keep archived narratives of various safety concerns related to multiple consumer products. They regulate and, if necessary, recall the products that pose severe safety concerns to consumers.

Saferproducts.gov has become an important database for the reporting of product-related safety

incidence in the United States. The website was authorized by Congress in 2008 and became active in 2011. Over the past few years, it has become progressively more accessible to the public. Based on this database, congressional testimonies have been presented by the Consumer Federation of America, Consumers Union, Kids in Danger (KID), Public Citizen, the US Public Interest Research Group (US PIRG), and others. These stakeholders have recommended that the CPSC merge additional data sources and resources into saferproducts.gov and increase data analysis efforts for the categories of harm and hazards that are listed in the database [24].

This study uses unstructured data to detect and categorize reports of safety concerns that may help identify products or product categories that the agency and manufacturers should be aware of because they are most likely to cause hazards to consumer safety. As unstructured textual data is difficult to analyze, particularly at great volumes, stakeholders and regulators may benefit greatly from monitoring real-time information.

2.2 Online safety surveillance literature

Online posts have emerged as a powerful new data source for firms and interest groups to mine for insights pertaining to product quality. The volume of online posts is enormous and expanding, allowing for more detailed and nuanced analyses. For surveillance of product quality and safety, online posts pertaining to consumers' experiences with products may be especially valuable. Given such an enormous volume of posts from which to draw, firms and interest groups have a rapidly updating data source that spans the range of consumer experiences with products. While social media and forum posts have been used to this end with some success [3, 4], online reviews represent a particularly targeted data source in which consumers have specifically written posts about their experience with products. Consumers detail their experiences with products and give manufacturers extensive details pertaining to product quality and performance [15].

Past studies that use online reviews as an indicator of product quality have focused on disentangling online reviews to reveal semantic trends [4, 12]. This work showed that online reviews that discuss product defects do not usually refer to strong emotions. Reviewers may write in a very factual tone and wish for their post to be seen by others as unbiased, and as such the wording of the defect reports may be less polarizing than that in other online reviews [19]. Similarly, the text of the online reviews may not always be emotive when it comes to explaining safety hazards, and hence, it becomes difficult for traditional text analysis, such as sentiment analysis, to effectively detect safety hazards [2-4]. For example, consumers post a wide range of negative reviews, but most of these negative comments are complaints about the product's quality (e.g., color, size, value, effectiveness, etc.), and few of these negative comments are concerns regarding safety hazards associated with the product.

A major limitation of prior works has been the narrowness of their scope. Due to the linguistic specificity of certain safety hazards, many prior analyses have only analyzed a single industry at a time [2, 3, 12, 19]. In this work, we uniquely apply a transfer learning approach to apply insights from a government dataset to a cross-industry sample. In addition, we utilize topic modeling approaches to shed further light on textual details such as the types of injury reports observed.

2.3 Topic modeling

Topic modeling is a statistical process used to analyze a corpus of text and delineate between distinctive clusters of words, or topics, that represent the major thematic emphasis of the corpus. For instance, the words "drive", "steering", "headlights", "brake", and "wheels" may pertain to a topic about cars. A corpus is generally comprised of multiple topics, and each document within that corpus may refer to a mixture of several of those topics together.

Most topic modeling is unsupervised, indicating that the techniques do not rely on training data or many rigorous assumptions about the underlying textual data [8]. Rather, these approaches instead are built upon the foundation that words that appear together in similar contexts also have related meanings [26]. Two of the most popular topic modeling techniques are Latent Semantic Analysis (LSA) [9] and Latent Dirichlet Allocation (LDA) [5]. LSA uses singular value decomposition to reduce the dimensionality of document-term matrices, revealing underling distributional linguistic patterns. LDA is a modification of probabilistic Latent Semantic Indexing (pLSI) [13] and uses a hierarchical Bayesian model to allocate words to topics. In most modern text mining research, LDA has emerged as the more popular development of the technique [8].

2.4. Transfer learning

Transfer learning is a machine learning process in which information obtained or learned from one domain is reapplied to another domain. Most machine learning processes assume some degree of homogeneity between the distribution and features of

training data and the distribution and features of future datasets to which machine learned models may be applied [22]. However, these assumptions may be tenuous in many real-world applications. For instance, datasets collected in the future may shift in its distribution and features. Even if the distribution and features of a particular domain were constant over time, it is often impossible to obtain sufficiently large quantities of training data from a domain of interest, and the process of generating training data may be very expensive and/or time-intensive. Transfer learning is often appropriate for these situations in which training data in one domain is insufficient to build a predictive model with high performance or is expensive to obtain or curate. Instead, there may be more data available from a related but distinct domain, and the insights garnered may apply to both domains [22]. This approach has become popular across numerous application areas in recent years [23, 27]. In this paper, we consider a case in which the target classification (safety hazards, or more specifically injury reports in amazon.com reviews) is particularly rare. Thus, the application of insights gained from the related domain of the CPSC's saferproducts.gov dataset may be a more practical approach to effectively analyze this problem.

3. Datasets and data coding

3.1. Saferproducts.gov dataset

The CPSC maintains saferproducts.gov, a repository that contains specific reports of product safety-related incidents. In Table 1, we present descriptive statistics on the injury types observed in this dataset. Some incidents reported to this site may be severe enough to result in injuries to consumers; however, in other cases, a hazardous scenario that represents the potential for injury is reported. The dataset contains a narrative describing each incident, a description of the product involved, its manufacturer, where the incident occurred, whether and to what extent an injury occurred, whether the product was damaged or modified before the incident, and additional information.

Just under one-third of reports indicate that an injury occurred; in many cases, however, the nature of these injuries is not specified in the report. While some reports are initiated by consumers, others are initiated by public safety entities, governmental bodies, health care professionals, and other interested parties. The database has been maintained since its inception in 2011, and as of 2019, it contains 39,613 records.

Table 1. Saferproducts.gov dataset injury report descriptive statistics.

Injury type	Count (percentage)
Injury	12,160 (30.7%)
First aid	*3,238 (8.2%)*
Emergency department	*1,514 (3.8%)*
Hospital admission	*593 (1.5%)*
Death	*137 (0.3%)*
Other or unspecified injury	*6,678 (16.9%)*
No injury	27,453 (69.3%)
Total	**39,613 (100.0%)**

3.2 Amazon.com dataset

We obtained a large dataset of product reviews posted on amazon.com, the world's largest e-commerce retailer [18]. To ensure a cross-industry sample, we chose 17 distinct product categories for inclusion in our analysis.

As our dataset was initially unlabeled (reviews were not marked for whether they referred to a safety hazard report or injury report), we performed this process manually. We recruited teams of undergraduate business students from a large public research university for manually coding (or "tagging") each review. Each team was assigned to a distinct product category. Each tagger was given a set of instructions describing the tagging assignment and was asked to tag about 200 reviews in a binary fashion: safety hazard or no safety hazard. A total of 124,289 reviews across the 17 industries were assigned to the taggers at random. Due to random assignment, there was some overlap in which multiple taggers tagged the same review; as such, 181,999 total tags were generated across the 124,289 reviews. Per the discussions in prior research [12], we reconciled any disagreements between taggers using a majority conservative decision rule: we used the majority vote of the taggers as the final label for each review. If the votes were tied, then we use the most conservative ("safety hazard") label. For these initial stages of our analysis, we sought to capture any possible safety hazard reports, and we would eliminate any false positives in later stages of verification.

Table 2. Initial amazon.com dataset safety hazard descriptive statistics.

Product category	Total tags	Unique reviews	Unique safety hazard tags (percentage)
Baby products	27,981	16,930	850 (5.0%)
Blenders	21,095	16,869	577 (3.4%)
Car seats	22,635	20,000	1,438 (7.2%)
Clothing	6,782	3,557	111 (3.1%)
Dishwashers	6,052	4,043	56 (1.4%)
Elderly products	15,259	14,443	617 (4.3%)
Furniture	7,643	6,674	48 (0.7%)
Garden tools	6,245	4,065	70 (1.7%)
Household products	7,705	3,612	104 (2.9%)
Musical instruments	1,605	1,405	3 (0.2%)
Office products	6,503	3,185	18 (0.6%)
Power tools	6,176	3,050	189 (6.2%)
Refrigerators	6,173	4,742	32 (0.7%)
Small appliances	11,450	5,519	97 (1.8%)
Smartphones	5,700	4,254	24 (0.6%)
Sports equipment	7,710	4,807	229 (4.8%)
Toys	15,285	7,134	475 (6.7%)
Total	**181,999**	**124,289**	**4,938 (4.0%)**

To ensure the reliability of the tagging process, we assigned a lead tagger to each project. The lead tagger tagged a random set of reviews for their project, overlapping with the other student tags on that project. If the lead tagger's tags and the other students' tags show high levels of agreement, then it suggests that the tagging was of high quality, and there were not substantial disagreements in the interpretation of the tagging assignment. We observed at least 84% agreement and Cohen's κ [7] values of at least 0.67 for each industry, reflecting "substantial agreement" per Landis and Koch [17] and "fair to good" agreement per Fleiss et al. [11]. Thus, the tagging protocol was applied consistently, and the resulting dataset is of high quality.

Across the 124,289 unique review analyzed in our dataset, 4,938, or 4.0%, were deemed to refer to safety hazards. However, the rate of safety hazard reports varied by industry from a low of 0.2% for musical instruments to a high of 7.2% for car seats. We detail descriptive statistics on our dataset in Table 2.

Having used a majority conservative decision rule to reconcile tagging, and recognizing the propensity of taggers to often over-tag the target classification [4, 12], we reverified the tags of all reviews that were tagged as safety hazards in our initial analysis. First, we distributed the 4,938 reviews to a final team of undergraduate students, asking these students also to determine whether the reviews referred to safety hazards. Of these 4,938 reviews, this team of taggers identified 1,389 as referring to true safety hazards. Second, as a final stage of validation, a team of graduate students carefully reviewed each of the 1,389 safety hazard-tagged review, verifying that 740 reviews referred to true safety hazards.

Table 3. Amazon.com dataset incident report descriptive statistics.

Incident type	Count (percentage)
Injury	95 (12.8%)
First aid	*9 (1.2%)*
Emergency department	*3 (0.4%)*
Hospital admission	*4 (0.5%)*
Death	*1 (0.1%)*
Other or unspecified injury	*78 (10.5%)*
No injury	645 (87.2%)

For comparison with the saferproducts.gov dataset, the graduate students also tagged the 740 verified amazon.com safety hazard reports for injury reports using the same taxonomy as in the former dataset. Descriptive statistics generated from this analysis are reported in Table 3. Like the saferproducts.gov dataset, most of the reports actually do not reference injuries, and an even smaller portion of the safety hazard reports on amazon.com do so. In the 95 cases that an injury was reported, the severity of the injury was other or unspecified in 78 cases (82.1% of the injuries). Due to the small number of true positives in the amazon.com dataset, it would be quite difficult to generate a

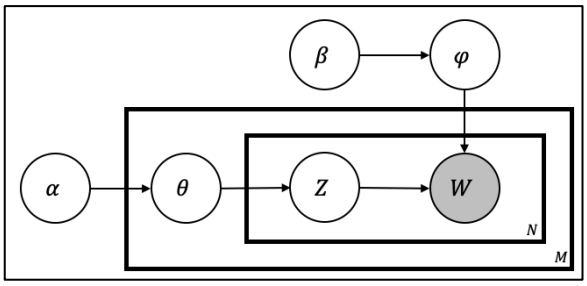
meaningful predictive model using it as a training set. Instead, transfer learning in which insights from the saferproducts.gov dataset are applied to this new domain may offer higher quality analyses.

4. Methodology

4.1. Topic modeling methods

For topic modeling, we utilize LDA, which is a multi-tiered hierarchical Bayesian model [5]. LDA is efficient even for situations in which the document-term matrix is sparse and/or in which the dataset is large and high-dimensional. Supposing that the dataset contains M documents and N words, let α represent the per-document Dirichlet parameter. In turn, β represents the per-topic Dirichlet parameter. Then, θ is the topic distribution for each document, and φ is the word distribution for each topic. Finally, let z represent the topic for the given word in the given document, and let w represent the word being analyzed. Then, Figure 1 shows the inner workings of LDA in graphical plate notation. The outer plate refers to the document-level analysis, while the inner plate refers to the word-level analysis within each document.

Figure 1. Graphical plate notation model of LDA (adapted from Blei et al. [5]).



For further details on the LDA methodology, we refer the reader to the initial study by Blei et al. [5]. In our study, we use LDA to generate topics for each of the injury classifications present in our saferproducts.gov dataset. By determining the latent topics present in these records, we may better understand the nature of injuries reported in consumer products as well as some of the top terms used to report such injuries. In the future, practitioners can use these results to categorize and prioritize their safety surveillance results as efficiently as possible.

4.2. Supervised smoke term generation

"Smoke terms" refer to distinctive words and/or phrases that are especially prevalent in records referring to the target classification, in this case product safety hazard reports [2-4, 12]. Various information retrieval approaches have been proposed for selecting appropriate candidate terms [2-4, 12]; however, the Correlation Coefficient (CC score) algorithm [10] has proved to be one of the most popular and high-performing. This technique was originally suggested by Ng et al. [20] and later expanded upon by Fan et al. [10]. Consider a corpus that contains a set of many documents, some of which are relevant (say, emergency department incident reports) and some of which are not. Furthermore, some of these documents include word *i*, and some of them do not. Table 4 defines the relationships between these document relevance and word inclusion (exclusion).

Table 4. Contingency table for each word's inclusion (exclusion) from each document (adapted from Fan et al. [10])

	Document is relevant	**Document is non-relevant**	**Row total**
Document contains word	A	B	$A + B$
Document does not contain word	C	D	$C + D$
Column total	$A + C$	$B + D$	N

The CC score method is based on the chi-square distribution and assesses the relevance of each word as follows in (1):

$$Relev = N \times (AD - CB) \times (A + B) \times (C + D) \quad (1)$$

Using this approach, a relevance score is generated for each word that appears in the training set, where higher relevance scores suggest words that occur very frequently in relevant documents (true positives) and very infrequently in irrelevant documents (true negatives). As such, these words may be meaningful predictors of relevant documents.

We partitioned our saferproducts.gov dataset into a training set (80%) and holdout set (20%) so that we could both generate smoke terms and evaluate their

performance in-domain before applying them to a separate domain. We then utilized the CC score algorithm to generate relevance scores for each term in the training set. After obtaining the highest-scoring terms, we removed stop words, common brand names, and common product categories. We then stored the 300 unique terms that corresponded with the highest relevance scores [2-4].

An analyzing a future dataset (transfer learning), we generate a "smoke score" for each record. To compute this smoke score for a given record, we find any occurrences of smoke terms in that record, each time incrementing our smoke score by the smoke term's relevance score as indicated by the CC score algorithm. To arrive at a final ranking, we simply sort the records in our dataset from highest to lowest smoke scores. The records with the highest smoke scores are deemed most likely to refer to the target classification. This smoke term approach is not interconnected with the aforementioned LDA approach; rather, the two approaches provide complementary information.

5. Results

We first ran LDA to determine the latent topics present in our saferproduct.gov dataset. We used the freely available Python implementation of LDA (see https://pypi.org/project/lda/) for this analysis. We ran the LDA analysis separately for each incident type (first aid, emergency department, and hospital admission). We were unable to run the analysis on reports of death, as we did not have enough records to perform the analysis. We experimented with numbers of topics ranging from 10 to 30; we found that 10 topics tended to yield the best human-interpretable results. We display the titles of our 10 LDA-generated topics in Table 5. We noticed some overlap in the topics generated between the different incident types. These topics are *italicized.*

Interestingly, although we observed considerable overlap in the topics generated between the different incident types, the specific words that comprised each topic varied in accordance with the severity of the incident type. The topics observed were generally consistent with the severity of the incident type. For instance, head/concussion and swallow injuries are among the most severe, and we only observed these topics for the hospital admission incident type.

In Table 6, we show the top words associated with an exemplar topic, heat/burns, across all three incident types analyzed. The intensity of the words appears to escalate as the incident types escalate, changing from words such as "hot" and "warm" for first aid to words such as "flame" and "smoke" for emergency department and finally words such as "explode" and "blaze" for hospital admission. These topics characterize the types of narratives that manufacturers and interest groups may expect to see around hazardous products. In addition, this analysis allows for the rapid delineation between the severity of these narratives.

Table 5. Titles of 10 LDA-generated topics for each incident type

First aid	Emergency department	Hospital admission
product name / ID	*product name / ID*	*product name / ID*
heat / burn	*heat / burn*	*heat / burn*
falling	*falling*	*falling*
child hazard	*child hazard*	head / concussion
cuts / laceration	*cuts / laceration*	swallow injury
contact seller	*contact seller*	mold / bacteria
rash / skin irritation	*eye / face injury*	*eye / face injury*
battery / electrical	*foot / ankle injury*	*foot / ankle injury*
defective	*hospital visit*	*hospital visit*
shattered glass	bandaging / treatment	hand / arm injury

Table 6. Top words in heat/burn topic across incident types.

Incident type	Top words
First aid	fire, hot, burn, warm, temperature, hand, start
Emergency department	fire, gas, burn, grill, degree, flame, smoke
Hospital admission	fire, burn, fuel, degree, explode, blaze, catch

Next, we used the CC score algorithm to generate candidate smoke terms for each incident type. After removing stop words, common brand names, and common product categories, we retained the top 300 highest scoring smoke terms for each incident type. In Table 7, we show the top smoke terms across the three incident types.

Table 7. Top smoke terms across incident types.

Incident type	Top words
First aid	finger, cut, hand, skin, sharp, fingers, bleeding, burns, rash, thumb
Emergency department	emergency, stitches, er, hospital, laceration, bone, pain, ambulance, treatment, rushed
Hospital admission	surgery, hospital, admit, icu, ambulance, fracture, surgeon, shatter, suffer, skull

The first aid smoke terms typically referred to small injuries, such as cut or burn injuries to hands or elsewhere on the skin. The emergency department smoke terms escalated, referring to trips to the hospital in an ambulance, stitches, and other medical treatments. Finally, the hospital admission smoke terms escalated further, referring to items such as surgeries or the intensive care unit (ICU).

Next, we tested the performance of the smoke terms. We assessed performance in two senses. First, we tested the performance of the smoke terms on the holdout set from the saferproducts.gov dataset (recall that we held out an unseen 20% of that dataset). Second, we attempted to transfer the knowledge garnered from the saferproducts.gov dataset, applying those smoke terms to detect mentions of injuries in the amazon.com dataset. We used the smoke terms and the associated relevance scores (weights) to rank all of the records in each of these sets from highest to lowest, where the highest ranked records were most likely to refer to true positives. Then, we can choose any arbitrary cutoff of the top N-ranked reviews (e.g., supposing that $N = 100$, we consider the top 100-ranked reviews) and examine the performance of the smoke terms within those records.

We assess performance according to four metrics. First, we calculate precision, or the proportion of the records identified within the cutoff that are actually true positives. For instance, if we are interested in the top 100-ranked reviews and observe 20 true positives within that cutoff, then precision is 20 / 100 or 0.200. Second, we calculate recall, or the proportion of all positive records that were identified within the cutoff. For instance, if we identified 20 true positives within our cutoff out of a possible 60 true positives in our dataset, then recall is 20 / 60 or 0.333. Generally, we observe an inverse relationship between precision and recall. At lower cutoffs, we might expect to observe high precision as the top-ranking records are the easiest to classify accurately, but they only represent a small portion of all true positives, so recall may be low. As the cutoff increases, more true positives are identified, but classification is more difficult, so the overall precision decreases as recall improves. The choice of the balance between these two criteria is a matter of some debate, but we present a range of options to practitioners so that a manager can choose an option that makes the most sense for their use case. Third, we calculate F-measure, which is a weighted compromise between precision and recall (specifically, the harmonic mean). Fourth, we calculate lift, or the ratio of the number of true positives identified within the cutoff to the number of true positives that one would expect to identify within that cutoff at a rate of random chance. For instance, if 60 true positives exist in the dataset out of 1000 records, then we would expect to observe 6 true positives in the top 100 reviews if we used random chance classification. If we actually identified 20 true positives within this cutoff, then lift is 20 / 6 or 3.333. We present precision, recall, F-measure, and lift values at cutoffs of the top 50-, 100-, 200-, 500-, and 1,000-ranked reviews for both datasets in Table 8.

The results from the saferproducts.gov holdout set indicate that high-performing smoke term lists were generated for each incident type. Performance was particularly strong for the hospital admission smoke term list, where the lift metric indicated that classification performance was as much as 27.648 times that of random chance. We observed the aforementioned relationship between precision and recall such that precision was particularly strong at lower cutoffs, and recall was particularly strong at higher cutoffs.

Classification in the amazon.com dataset to which we wished to transfer information was especially difficult. This dataset consisted of 124,289 reviews, of which just 9 (0.007%) referred to first aid, 3 (0.002%) referred to emergency department, and 4 (0.003%) referred to hospital admission. Thus, when classifying by random chance, one would expect to have to read thousands of online reviews before identifying any such reviews. Using our smoke term lists, however, we found that the transfer of information from the saferproducts.gov dataset was remarkably successful. The precision metrics appear low because the target classification was so rare, but likewise the recall metrics are considerable, and the lift metrics indicate that performance was generally hundreds of times better than would be expected with random chance classification. Thus, although the target classification is incredibly rare, the application of information garnered from the saferproducts.gov dataset makes prioritization of this content possible.

Table 8. Smoke term performance in holdout set and unseen transfer dataset across incident types.

	Cutoff	Precision / recall / F-measure / lift		
		First aid	Emergency department	Hospital admission
Saferproducts.gov holdout set	50	0.220 / 0.017 / 0.032 / 2.556	0.420 / 0.074 / 0.126 / 11.398	0.360 / 0.180 / 0.240 / 27.648
	100	0.190 / 0.029 / 0.050 / 2.208	0.470 / 0.166 / 0.245 / 12.755	0.300 / 0.300 / 0.300 / 23.040
	200	0.235 / 0.071 / 0.109 / 2.730	0.430 / 0.304 / 0.356 / 11.669	0.245 / 0.490 / 0.327 / 18.816
	500	0.262 / 0.198 / 0.226 / 3.044	0.332 / 0.587 / 0.424 / 9.010	0.154 / 0.770 / 0.257 / 11.827
	1,000	0.263 / 0.398 / 0.317 / 3.056	0.222 / 0.784 / 0.346 / 6.025	0.087 / 0.870 / 0.158 / 6.682
Amazon.com dataset	50	0.040 / 0.222 / 0.068 / 552.396	0.020 / 0.333 / 0.038 / 828.593	0.020 / 0.250 / 0.037 / 621.445
	100	0.040 / 0.444 / 0.073 / 552.396	0.020 / 0.667 / 0.039 / 828.593	0.020 / 0.500 / 0.037 / 621.445
	200	0.025 / 0.556 / 0.048 / 345.247	0.010 / 0.667 / 0.020 / 414.297	0.010 / 0.500 / 0.020 / 310.723
	500	0.012 / 0.667 / 0.024 / 165.719	0.006 / 1.000 / 0.012 / 248.578	0.006 / 0.750 / 0.012 / 186.434
	1,000	0.006 / 0.667 / 0.012 / 82.859	0.003 / 1.000 / 0.006 / 124.289	0.003 / 0.750 / 0.006 / 93.217

6. Conclusion

In this paper, we utilized topic modeling and transfer learning techniques to improve safety monitoring techniques pertaining to incidents in which consumers were injured by products. Our topic modeling of the saferproducts.gov dataset using LDA revealed latent topics in first aid, emergency department, and hospital admission incidents. While there was some overlap in topics among these incident types, we observed a difference in terms suggesting an escalation in the language used in narratives. We generated supervised smoke terms for each of these incident types, finding that these terms worked well not only for their in-domain holdout set but also when applied to a new domain of amazon.com reviews.

While prior works have assessed identifying reports of safety hazards in online posts [2-4, 12], our work is unique in that we focused largely on injury reports in these reviews. These records are of particularly high value to both firms and interest groups, as they represent pressing issues in need of the most immediate solution. Identifying these possible issues as quickly as possible allows firms to remediate, avoiding possible financial and legal issues associated with product recalls. Furthermore, our work is unique in that we applied our technique in a cross-industry setting. We examined a saferproducts.gov dataset that spans all consumer products as well as 17 unique product categories from amazon.com. While supervised smoke terms have generally only been effective in the context of a singular industry, we found that there is great utility in the application of these techniques to a multi-industry context.

Our work is subject to several important limitations. Our analysis of the amazon.com dataset required an enormous effort of manual tagging, which involves some subjectivity on the part of taggers. We took steps to reduce the effect of this subjectivity by checking for agreement between taggers and performing several iterative rounds of tagging. A further limitation is that our work in this paper was limited to a range of consumer products. Future research may explore extensions of our work in which these techniques are applied to additional industries, such as those with industrial or workplace safety implications. A further limitation is that, while the machine learning and transfer learning techniques implemented in this paper performed well, alternative techniques are also available. Future research may explore the performance of these alternatives relative to this paper's techniques.

References

[1] Spectrum brands ordered to pay civil penalty for failure to report and post-recall sales of defective SpaceMaker coffee carafes, in: Department of Justice, (2017).

[2] A.S. Abrahams, W. Fan, G.A. Wang, Z. Zhang, J. Jiao, An integrated text analytic framework for product defect

discovery, Production and Operations Management, 24(6) (2015) 975-990.

[3] 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.

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

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

[6] Y. Chen, S. Ganesan, Y. Liu, Does a firm's product-recall strategy affect its financial value? An examination of strategic alternatives during product-harm crises, Journal of Marketing, 73(6) (2009) 214-226.

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

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

[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6) (1990) 391-407.

[10] 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.

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

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

[13] T. Hofmann, Probabilistic latent semantic indexing, in: ACM SIGIR Forum, (ACM, 2017), pp. 211-218.

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

[15] N. Hu, P.A. Pavlou, J. Zhang, Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication, in: Proceedings of the 7th ACM Conference on Electronic Commerce, (ACM, 2006), pp. 324-330.

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

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

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

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

[20] H.T. Ng, W.B. Goh, K.L. Low, Feature selection, perceptron learning, and a usability case study for text categorization, in: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1997), pp. 67-73.

[21] J.Z. Ni, B.B. Flynn, F.R. Jacobs, Impact of product recall announcements on retailers′ financial value, International Journal of Production Economics, 153(2014) 309-322.

[22] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering, 22(10) (2009) 1345-1359.

[23] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, F. Provost, Machine learning for targeted display advertising: Transfer learning in action, Machine Learning, 95(1) (2014) 103-127.

[24] M.S. Robinson, A public health and data crisis you can help solve: CPSC's critical need for NASPGHAN's data, Journal of Pediatric Gastroenterology and Nutrition, 65(2) (2017) 133-134.

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

[26] P.D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics, Journal of Artificial Intelligence Research, 37(2010) 141-188.

[27] L. Xia, D.M. Goldberg, S. Hong, P.K. Garvey, Transfer learning in knowledge-intensive tasks: A test in healthcare text analytics, in: 25th Americas Comference on Information Systems, (Cancun, 2019).