Risk Analysis, Vol. 00, No. 0, 2020 DOI: 10.1111/risa.13651

Text Mining Approaches for Postmarket Food Safety Surveillance Using Online Media

David M. Goldberg[1,*] **Samee Khan**[2] **Nohel Zaman**[3] **Richard J. Gruss**[4] **and Alan S. Abrahams**[2]

Food contamination and food poisoning pose enormous risks to consumers across the world. As discussions of consumer experiences have spread through online media, we propose the use of text mining to rapidly screen online media for mentions of food safety hazards. We compile a large data set of labeled consumer posts spanning two major websites. Utilizing text mining and supervised machine learning, we identify unique words and phrases in online posts that identify consumers' interactions with hazardous food products. We compare our methods to traditional sentiment-based text mining. We assess performance in a high-volume setting, utilizing a data set of over 4 million online reviews. Our methods were 77–90% accurate in top-ranking reviews, while sentiment analysis was just 11–26% accurate. Moreover, we aggregate review-level results to make product-level risk assessments. A panel of 21 food safety experts assessed our model's hazard-flagged products to exhibit substantially higher risk than baseline products. We suggest the use of these tools to profile food items and assess risk, building a postmarket decision support system to identify hazardous food products. Our research contributes to the literature and practice by providing practical and inexpensive means for rapidly monitoring food safety in real time.

KEY WORDS: Food safety; online reviews; text mining

1. INTRODUCTION

The globalization of the food industry in recent years has resulted in increased challenges for firms as they attempt to navigate food safety concerns (Garre, Boué, Fernández, Membré, & Egea, 2019; Njage et al., 2019). Contaminated food items that subject consumers to risk of foodborne illnesses are quite nefarious, and they are also difficult for firms to monitor effectively. Firms at all levels of the supply chain, including producers, distributors, and retailers, must grapple with concerns stemming from food safety. According to the Centers for Disease Control (CDC), an estimated 48 million cases of foodborne illness are contracted in the United States every year, which cause about 128,000 hospitalizations and 3,000 deaths (Scallan, Griffin, Angulo, Tauxe, & Hoekstra, 2011). Thirty-one so-called "major pathogens" cause an estimated 9.4 million of the 48 million annual cases of foodborne illness, but the remaining 38.6 million cases, a vast majority, are of unknown origin (Scallan et al., 2011). Some cases develop into major public new stories, which result in alerts for both consumers and firms. For example, a 2019 incident involving ground beef in the United States resulted in a major public alert across 10 states as a total of 209 people were infected with *Escherichia Coli*, causing 29 hospitalizations and the recall of nearly 200,000 pounds of ground beef (Outbreak of E. Coli Infections Linked

[1]San Diego State University, San Diego, CA, USA.
[2]Virginia Tech, Blacksburg, VA, USA.
[3]Loyola Marymount University, Los Angeles, CA, USA.
[4]Radford University, Radford, VA, USA.
*Address correspondence to David M. Goldberg, 5500 Campanile Drive, Mail Code 8234, San Diego, CA 92182, USA; tel: +1-(619)-594-0341; dgoldberg@sdsu.edu

to Ground Beef, 2019). However, these major public incidents represent only a small fraction of the foodborne illness cases that occur in the United States each year, as most cases are unreported (according to CDC numbers, about 1 in 1,866 cases, or 0.05% are reported (Tack et al., 2019)), and even those that are reported are often difficult to effectively track to their origin. In addition, food-related recalls in the United States have only increased in frequency in recent years, rising by 10% between 2013 and 2018, which many experts associate with the increasing complexity and globalization of food supply chains (Ducharme, 2019). The problem has been especially pronounced for meat and poultry, for which recalls increased by 83% over the same period (Ducharme, 2019). Some experts associate this large jump with recent regulatory relaxations in premarket inspection protocol in these industries (Ducharme, 2019). Given these concerns, rapid monitoring of product safety in the food industry is vital.

Management and monitoring of product safety is made especially difficult in the food industry for three major reasons. First, the nature of the food industry leads to inevitable variations in quality (Mokhtari & Van Doren, 2019; Trienekens & Zuurbier, 2008). Numerous producers are involved in most commercial food supply chains, resulting in great disparities in quality. While some differences in quality are controllable by improving production practices, other factors such as weather/climatic conditions, seasonality, and biological factors are generally outside of producers' control (Trienekens & Zuurbier, 2008). Second, quality inspections in the food industry are often ineffective at identifying potential problems. Yao and Parlar (2019) commented that many products may pass internal or external quality inspections, and they can still result in serious safety concerns leading to product recalls after reaching the market. As food products are processed in a variety of environments before reaching consumers, an issue in any one of these environments could cause initial testing to yield misleading results. Third, a lack of traceability in food supply chains makes potential problems difficult for firms to effectively locate and manage. Kshetri (2018) argues that the inherent complexity of food supply chains makes record-keeping unwieldly and challenging; thus, most firms are unable to monitor the status of other firms in the same supply chain. Some recent research has suggested the use of blockchain technology to address this problem (Kamilaris, Fonts, & Prenafeta-Boldύ, 2019), although such solutions have not yet achieved mainstream adoption (Behnke & Janssen, 2020; Kamilaris et al., 2019). Even after a problem is identified, firms struggle to identify the cause, particularly as the problematic food may have proliferated through the supply chain (Mokhtari & Van Doren, 2019). Because previously, safe food products may become unsafe at any stage of the supply chain, postmarket monitoring is crucial for all firms taking part in food supply chains. For example, in 2015, Chipotle Mexican Grill restaurants suffered greatly after some of their food was contaminated with *E. Coli*, causing illness in many customers (Kshetri, 2018). As Chipotle attempted to identify the cause of the contamination, the extensive shutdowns of restaurants and subsequent investigations were time-consuming and costly to its business stature.

The major regulatory bodies for food products in the United States are the United States Department of Agriculture Food Safety and Inspection Service (USDA FSIS) and the Food and Drug Administration (FDA). These agencies detect unsafe products in four major ways: a firm reports that a potentially hazardous food item is on the market; a governmental random sampling program reveals an issue with a food product; a governmental field inspector discovers an unsafe food product onsite; or another public health agency reports epidemiological data that suggest the presence of unsafe food items (FSIS Food Recalls, 2015). Despite these best intentions of both firms and regulatory agencies, these means of detection suggest the potential for major contaminants to slip through the cracks. As prior literature has noted, food producers often struggle to identify and locate food contaminants in their supply chains, which results in dangerously slow reporting (Kshetri, 2018; Trienekens & Zuurbier, 2008). Given the scale of the global food industry, regulatory and public health agencies lack the means to thoroughly monitor all food products in real time, meaning that sampling and inspections will only capture a small fraction of possible foodborne contaminants. As such, current monitoring techniques expose firms and consumers to substantial risks. Some recent works have called for applications of text mining to food safety, as the ability to rapidly screen great volumes of textual data could help to identify potential food safety issues (Duggirala et al., 2015; Kate, Chaudhari, Prapanca, & Kalagnanam, 2014; Tao, Yang, & Feng, 2020). Most applications of text mining at regulatory agencies have been targeted at health reports pertaining to

medications rather than food safety and are considered developmental rather than routine (Duggirala et al., 2015).

In this work, we implement cutting-edge text mining techniques to utilize online media as a source of potential information for postmarket surveillance in the food industry. We propose a new food safety monitoring system (FSMS) utilizing text mining for sorting and prioritization of these comments. To do so, we gather a large data set of consumer reviews from Amazon.com in which consumers post about their experiences with products. In addition, we supplement this data set with data from IWasPoisoned.com, a particularly targeted data set on which consumers alert others to cases of food poisoning. We design a semiautomated machine learning methodology to discover the food safety hazard reports mentioned in consumer feedback. After utilizing our tool, firms can work with an automated shortlist of safety concerns. These shortlists allow firms to assess the need to inspect products further for contaminants, remediate products (for example, correct a packaging label that has omitted an allergen), or adapt production processes, suppliers, or ingredients to mitigate risks. In addition to applications for firms, regulatory agencies could use our tool to monitor industry, and consumers could use our tool to make safety-conscious decisions about their food intake.

Our research has valuable contributions to the literature in risk analysis, where food safety is a considerable concern due to its scale and complexity. First, the integration of our semiautomated text mining approach into current food safety surveillance approaches would benefit stakeholders including consumers, firms, and regulatory agencies as they attempt to rapidly pinpoint major risk factors. Given the scale and complexity of modern supply chains, unifying the literature in text mining for safety surveillance (Abrahams, Fan, Wang, Zhang, & Jiao, 2015; Abrahams, Jiao, Wang, & Fan, 2012; Adams, Gruss, & Abrahams, 2017; Goldberg & Abrahams, 2018) with pressing modern food safety concerns has the potential to improve social media analytics for global food supply chains. Several recent works have called for the application of data analytics and text mining to food safety (Duggirala et al., 2015; Kate et al., 2014), but works in this area have generally been retrospectives on social media use during food safety recalls (Tse, Loh, Ding, & Zhang, 2018) as opposed to predictive. Thus, our text mining approach represents a major advancement in this area of study. Second, while review-level analytics have been explored in prior work (Abrahams, Fan, et al., 2015; Abrahams, Jiao, et al., 2012; Goldberg & Abrahams, 2018; Zaman, Goldberg, Abrahams, & Essig, 2020), we extend these methodologies to provide novel analytics that use text mining to aggregate risk on the product level, considering factors such as product circulation and consumer concurrence (multiple hazard reports for the same product). We apply these analyses in the food industry, though our work has implications for real-time safety surveillance across industries. Third, we validate our findings by comparing our model's high-risk designations to the opinions of a panel of highly experienced food safety experts. We show substantial concordance between our model's outputs and expert opinion. In doing so, we show the external validity of text mining online reviews for safety surveillance, which demonstrates the value of surveilling online reviews in the food industry and provides the first rigorous expert validation of this practice, which has been suggested in additional industries (Abrahams, Fan, et al., 2015; Adams et al., 2017; Goldberg & Abrahams, 2018).

2. LITERATURE REVIEW

2.1. Food Production and Food Safety

Folkerts and Koehorst (1998) define a food supply chain as a "set of interdependent companies that work closely together to manage the flow of goods and services along the value-added chain of agricultural and food products, in order to realize superior customer value at the lowest possible costs." Van Der Vorst, Tromp, and Zee (2009) argue that there are two broad categories of food supply chains: those that produce fresh agricultural products and those that produce processed food products. Fresh agricultural products, such as fresh fruit and vegetables, involve a wide array of growers to produce sufficient food to meet demand in addition to further providers that handle packing, storage, and transportation before retailers finally sell these products to consumers. Processed food products, such as packaged meats or canned food, involve many of the same supply chain steps as fresh agricultural products in addition to an added step of food processing in which raw materials (food) are altered, usually preserving their shelf life. This step adds additional complexity to the supply chain, although once packaged, food products generally have more consistent quality. Both forms of supply chains

have become increasingly complex in recent years as global demand for food has increased; Staniškis (2012) argued that, due to global population growth, demand for food is approximately 30% higher than the planet's natural capacity. As such, retailers have been forced to utilize numerous growers to meet demand, increasing the complexity of supply chains and raising questions as to the quality of production.

Food safety risks are often understood along two axes: likelihood (or frequency) and severity (or consequence). This model has been common in the literature (Li, Bao, & Wu, 2018) and has also been adopted by major safety organizations, such as the Food and Agriculture Organization of the United Nations (FAO, 2017) and the European Food Safety Authority (Chatzopoulou, Eriksson, & Eriksson, 2020). Hazard Analysis and Critical Control Points (HACCP) is an internationally recognized food risk management system originated in the United States, which is also built upon this foundation (Wallace, Holyoak, Powell, & Dykes, 2014). Much past work has focused on food safety from a food science perspective, such as identifying biological agents whose effects are especially nefarious (Cox Jr, Popken, Sun, Liao, & Fang, 2020; Leuschner et al., 2010). In recent years, researchers and industry practitioners have sought to address some food safety concerns with technologies such as enterprise supply chain databases (LeBlanc et al., 2015) and blockchain technology (Behnke & Janssen, 2020; Kamilaris et al., 2019). Tao et al. (2020)'s recent literature survey suggests that the use of data analytics and text mining to monitor food safety is promising, but the area has received little study. Nsoesie, Kluberg, and Brownstein (2014)'s analysis of online reviews suggests that they are a potentially fertile data source for postmarket surveillance.

2.2. Text Mining

Text mining refers to the process of parsing text (such as online social media posts, forum posts, or product reviews) using automated methods to derive valuable information. Text classification is a particular form of text mining in which text is sorted into desired categories. For instance, in our study, we might wish to distinguish safety hazard-related text from unrelated text. Supervised machine learning is a popular approach for text classification, which involves using training data to build a predictive model (Abrahams, Jiao, Fan, Wang, & Zhang, 2013; Brahma, Goldberg, Zaman, & Aloiso, 2020; Goldberg & Abrahams, 2018). This technique requires a preexisting data set in which text is labeled to represent the desired classifications. A random selection of text records is obtained, and each record is manually examined and labeled to establish a "ground truth" for classification. Text is partitioned into a training set, which is used to build a high-performing predictive model, and a holdout set, which is initially unseen by the model and thereafter is used to validate the model's performance.

Sentiment analysis is a prominent text classification method that is useful for computing the direction and/or degree of the emotive content in text. For example, the phrase "the chili was awesome" is emotively positive, and the phrase "the salsa was terrible" is emotively negative. Sentiment analysis is utilized in a broad spectrum of text mining applications. For instance, sentiment analysis has been used to monitor consumer complaints in online forums (Hsieh, Ku, Wu, & Chou, 2012) and to identify tweets relevant to supply chain management issues (Singh, Shukla, & Mishra, 2018). However, conventional sentiment analysis constitutes limitations, particularly if the text comprises domain-related discussion not prevalent in positive/negative emotive valence. For example, Abrahams, Jiao, et al. (2012) evaluated sentiment analysis for rapidly discovering vehicle defects discussions in online fora, and the researchers found that sentiment analysis was not efficient when implemented on its own. For example, the most distinctive keywords related to vehicle defects, such as the distinct word "airbag," are nonemotive keywords that are pertinent only within the context of the specific industry, and sentiment was not always positively associated with product defects (Abrahams, Jiao, et al., 2012).

2.3. Online Safety Surveillance

In past few years, several papers have used supervised learning techniques as a means of discovering product safety hazards in online media (Abrahams, Fan, et al., 2015; Abrahams, Jiao, Fan, et al., 2013; Abrahams, Jiao, et al., 2012; Goldberg & Abrahams, 2018). Online reviews are a particularly popular data source, as they offer targeted feedback on product performance. Mummalaneni, Gruss, Goldberg, Ehsani, and Abrahams (2018) noted that the textual content of online reviews describing safety hazards is often factual rather than emotively strong. Hence, may be cumbersome for conventional text classification methods to detect safety hazards.

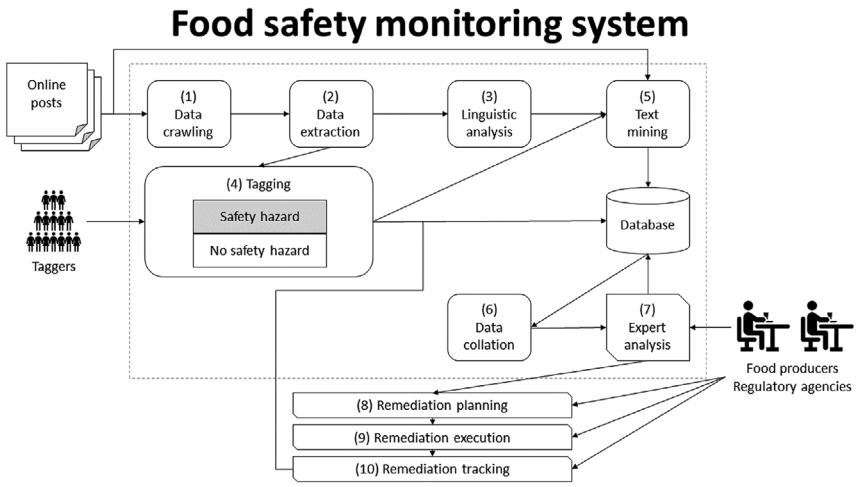

Fig 1. Food safety monitoring system (FSMS).

Recent research has only scratched the surface in using text mining for safety surveillance in the food industry. Song, Guo, Hunt, and Zhuang (2020) performed a descriptive analysis of the safety-related words and phrases used in blog posts concerning take-away food; however, as Tao et al. (2020) argue, there is a substantial need to develop new predictive tools that rapidly classify reports of unsafe foods.

Supervised text mining methods rely a prelabeled data set to train. Acquiring a labeled data set may be time-consuming and costly, but it can generate unique and robust results corresponding to a specific domain (Keyvanpour & Imani, 2013). Several past studies in text mining created dictionaries consisting of domain-related keywords for a specific target class (Abrahams, Fan, et al., 2015; Abrahams, Jiao, et al., 2012; Adams et al., 2017; Goldberg & Abrahams, 2018; Mummalaneni et al., 2018). These dictionaries contain key terms that are known as "smoke terms," which are significantly more prevalent in particular topics of interest in text documents. Smoke term methods are tuned to learn and extract the semantic properties of specific domain of interest, and hence, they may contain both emotive and nonemotive terms.

3. A TEXT MINING FRAMEWORK FOR FOOD SAFETY

We propose an FSMS for the rapid prioritization and identification of safety hazards in food items using online media. The proposed FSMS is semiautomated, meaning that many of the key aspects are performed extremely quickly with computer algorithms, but some final manual review is required from experts, particularly for determining how best to act upon identified safety hazards. We note that although we suggest that the FSMS is a useful aid to improve and augment current monitoring processes, it should not be the only method used to monitor food safety; food producers and regulatory agencies ought to continue quality testing and reporting in addition to the proposed FSMS. In Figure 1, we provide a graphical overview of the proposed FSMS.

Food quality management requires safety hazard identification (steps 1–7) as well as remediation (steps 8–10). The FSMS that we describe offers decision support for safety hazard identification in steps 1–7, which we discuss in this article; remediation processes performed by food producers and regulatory agencies are outside the scope of this work.

The quality management process begins with the crawling of online data to aggregate customer discourse (step 1). For this article, we focus our analysis on online posts from both Amazon.com and IWasPoisoned.com. Next, data extraction (step 2) is performed: metadata and actual text are extracted from the initial crawled data. This step delineates the text of the online post from additional fields, such as the date of the post, the author's name, and the title of the post. Next, linguistic analysis is performed (step 3), in which words are disambiguated. This process involves tokenization, the process of delineating where each word starts and ends. Practitioners may also use stemming or lemmatization to condense similar words (for example, "hurt," "hurts," and "hurting" are all forms of the same root word, "hurt"). Practitioners may also remove "stop words:" common English words like "the," "and," or "but" that are so common that they are unlikely

to be predictive. Tagging (step 4) involves manually labeling a large set of online reviews. In our study, reviews are labeled as "safety hazard" or "no safety hazard." This step establishes a standard for identifying specific words and phrases associated with online mentions of safety hazards. Once the reviews have been tagged, automated text mining (step 5) can be utilized to construct predictive modeling for rapidly sorting and prioritizing online reviews. All automated and manual text markups are stored in a database to allow for a historical record of safety hazard reports. Collation (step 6) of reviews involves sorting and prioritizing them based on the predictive modeling in preparation for analysis by an expert (step 7), who decides upon next steps.

Once the expert has performed their analysis, they may use it to inform remediation planning (step 8), in which a desired remediation strategy is chosen. Depending on the nature of the hazard, potential resolutions could include improving production facilities to mitigate contaminants, implementing heightened sanitation practices, choosing higher-quality ingredients, reevaluating ingredient selections, including warning labels on packaging, or recalling the product from the market. For instance, if the hazard is due to a bacterial contamination, then sanitation practices may be an important remediation strategy; on the other hand, if a product contains undeclared allergens, then producers must either revise their packaging or investigate alternative ingredients. In remediation execution (step 9), the strategy is implemented, and then the progress of the remediation is tracked (step 10). As safety hazards are rectified, this information is updated in the database to reflect successful resolution. Then, food producers and regulatory agencies can refocus on any additional areas of concern.

4. METHODOLOGY

4.1. Data Source and Data Coding

We utilized online reviews from Amazon.com, the world's largest e-commerce retailer, for our data set in this study (Ni, Li, & McAuley, 2019).[1] A web crawler collected this data set in 2018, and it includes all reviews of the selected products that had been posted on Amazon.com as of that time; thus, the reviews span an approximately 18-year period between 2000 and 2018. To ensure the legitimacy of potential safety hazard reports, we only consider "verified purchase" reviews, or reviews posted by Amazon.com accounts that had purchased the product that they were reviewing. We consider the reviews available in the "grocery and canned food" product category. Of the 5,074,160 total reviews available in this product category, we consider 4,437,360 "verified purchase" reviews, or 87.5% of all reviews available. We utilized 11,190 randomly drawn "verified purchase" reviews. These reviews discuss a great deal of topics, including the food's appearance, its taste, its value, and potential health/safety effects. As the data set was initially not labeled for its references to safety hazards, we first devised a coding scheme to guide our manual tagging process. The USDA FSIS delineates between three major classes of recall reflecting the risk level of food products as follows (FSIS Food Recalls, 2015):

- *"Class I - A Class I recall involves a health hazard situation in which there is a reasonable probability that eating the food will cause health problems or death.*
- *Class II - A Class II recall involves a potential health hazard situation in which there is a remote probability of adverse health consequences from eating the food.*
- *Class III - A Class III recall involves a situation in which eating the food will not cause adverse health consequences.*"

Thus, based on this scheme, we developed the following tagging instructions for taggers to classify each review as 1. "safety hazard" or 2. "no safety hazard."

(1) **Safety hazard**: We define a review to refer to a safety hazard when it indicates that the product may cause adverse health consequences. Examples included stomach discomfort, vomiting, diarrhea, and in some cases requiring medical treatment by a doctor or hospital. This classification includes USDA FSIS classes I and II such that there is at least some remote probability of the food product resulting in adverse health consequences.

(2) **No safety hazard**: We define a review to not refer to a safety hazard when it does not meet the qualifications mentioned above. The review may mention positive, neutral, or negative aspects of the product such its advertising, packaging, or quality (including taste or texture or other consumer expectations).

[1]This dataset is publicly available at https://nijianmo.github.io/amazon/.

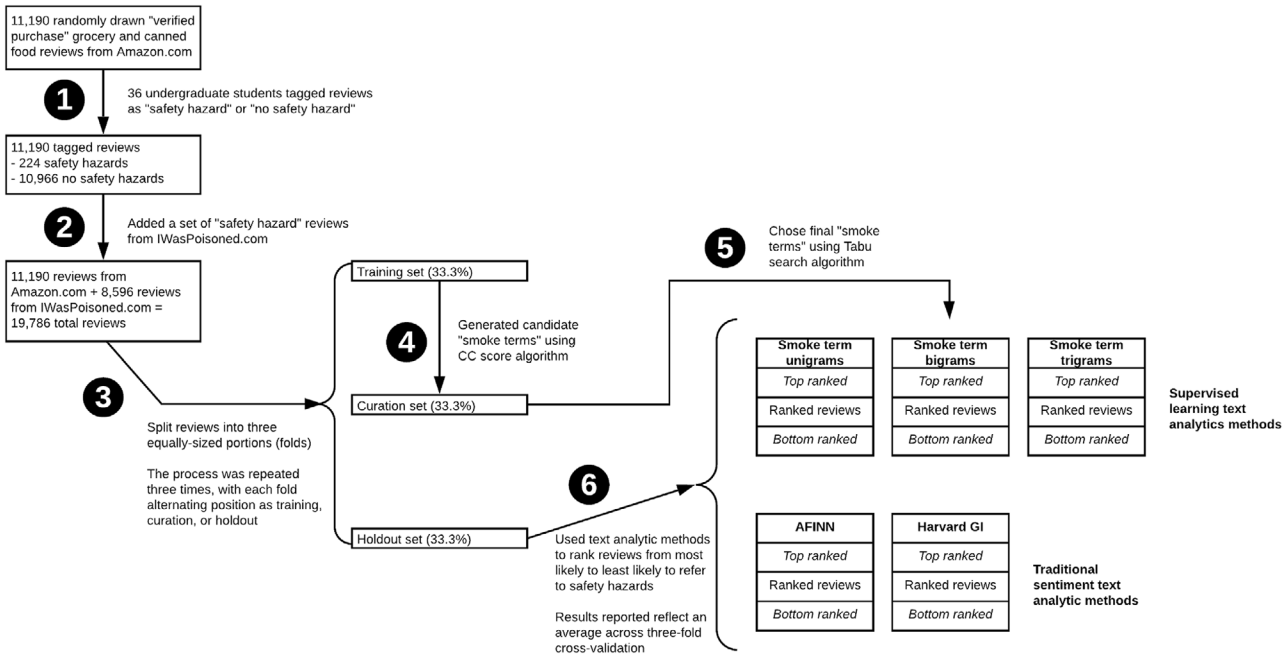

Fig 2. Overview of methodology.

4.2. Data Processing

Next, we elaborate each procedure that we executed for the data processing. Fig. 2 describes the steps that we implemented in for data processing in this study. Each numbered step in Fig. 2 is detailed in the corresponding numbered paragraphs below.

(1) Using the tagging protocol described above, 37 undergraduate business students at a large public research university manually tagged a random set of reviews from the pool of 11,190. The students were assigned the reviews at random, resulting in multiple tags for some reviews. We found that the students generated a total of 15,672 tags across the 11,190 unique reviews. To ensure the quality of the tags, we first checked the tags for internal consistency, comparing the student tags to one another in overlapping cases in which multiple students had assessed the same review. We observed excellent agreement, as student taggers selected the same tag 97.2% of the time. We also computed Cohen's κ (1968), a metric that compares this level of agreement to random chance. We observed a Cohen's κ score of 0.94, which Landis and Koch (1977) rate as "almost perfect agreement" and Fleiss, Levin, and Paik (2013) rate as "excellent agreement." To further assess the tagging quality against an external agent, a professional food scientist with a doctoral degree specializing in food safety tagged an overlapping set of 200 reviews. The professional food scientist's tags were then compared to the students' tags. We observed 95.0% agreement with a Cohen's κ score of 0.90, which Landis and Koch (1977) rate as "almost perfect agreement" and Fleiss et al. (2013) rate as "excellent agreement." Thus, we are convinced that the data set was tagged with high quality and in a consistent manner. Finally, we considered the few cases in which the student taggers disagreed with one another. In these cases, we reconciled the students' tags using a majority conservative decision rule. That is, we used a majority vote to decide upon the final decision where possible; if the taggers were tied, then we always chose the "safety hazard" class as the prevailing class (Goldberg & Abrahams, 2018; Law, Gruss, & Abrahams, 2017; Mummalaneni et al., 2018). From this data set, we identified a total of 224 safety hazard reports (2.0% of reviews) and 10,966 no safety hazard reports (98.0% of reviews).

(2) As the vast majority of the Amazon.com reviews did not refer to safety hazards, we needed further safety hazard reports to train

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

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

while not (stopping_condition()) // 200 iterations without improvement

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

    for i = 1 to length

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

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

    end for

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

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

end while

return best_solution
```

Fig 3. Pseudocode for Tabu search approach (adapted from Goldberg & Abrahams, 2018).

a meaningful predictive text mining model. Thus, we collected further set of textual data from a second website, IWasPoisoned.com. This website allows individuals to post reviews of food products that have caused illnesses as warnings to fellow consumers, firms, and regulators. Each IWasPoisoned.com review is curated both technically and by manual examination to mitigate inauthentic posts.[2] From this second data source, we added a random set of 8,596 reviews. We labeled each of these reviews as "safety hazards." We added this data set to the 11,190 tagged reviews from Amazon.com. As a result, we collected a total sample of 19,786 text documents (8,820 safety hazards, or 44.6%; 10,966 no safety hazards, or 55.4%) to perform further analysis.

(3) We divided our data set into three equally sized partitions: a training set, a curation set, a holdout set. The initial training set is used to determine possible candidate smoke terms for predicting safety hazards; the curation set is used to narrow down the candidate smoke terms to those that provide the best prediction; finally, the validation set is used to assess the efficacy of the final technique. To improve the robustness of our analysis, we performed this step using k-fold cross-validation ($k = 3$), wherein the analysis was repeated multiple times with each partition or fold alternating its position as training, curation, or holdout set (three folds) (Delen & Zolbanin, 2018). We ultimately report an average of the results observed across each analysis.

(4) Using our training set, we applied an information retrieval technique proposed by Fan, Gordon, and Pathak (2005) to generate an initial set of candidate smoke terms. The technique, known as the CC score algorithm, assigns each term in the training set a "relevance score," where higher relevance scores are thought to

[2]IWasPoisoned.com discusses this policy on their frequently asked questions page at https://iwaspoisoned.com/page/faq.

be more predictive. Let A equal the number of safety hazard-tagged reviews in which term i occurs; let B equal the number of no safety hazard-tagged reviews in which term i occurs; let C equal the number of safety hazard-tagged reviews in which term i does not occur; and let D equal the number of no safety hazard-tagged reviews in which term i does not occur. Finally, let N equal the total number of reviews, or $A + B + C + D$. Based on the chi-square distribution, the CC score defines the relevance of term i as:

$$relevance = \frac{\sqrt{N} \times (AD - CB)}{\sqrt{(A + B) \times (C + D)}}. \quad (1)$$

For example, suppose that there are 6,596 text documents in the training set (N), where 2,940 are safety hazard-tagged and 3,656 are no safety hazard-tagged. Consider the term "hurts." Suppose that this word occurs in 200 safety hazard-tagged text documents (A) and 100 no safety hazard-tagged text documents (B). It then does not occur in the remaining 2,740 safety hazard-tagged text documents (C) and 3,556 no safety hazard-tagged text documents (D). Thus, the relevance score for the word "hurts" would be 25,836. We retain relevance scores for each term for use as weights in later steps. Terms that occur very frequently in the target classification (safety hazards) and very infrequently otherwise receive particularly high scores. We used this technique to generate relevance scores for unigrams (one word), bigrams (two-word phrases), and trigrams (three-word phrases).

(5) Using our curation set, we applied the Tabu search approach suggested by Goldberg and Abrahams (2018) for fine-tuning the candidate smoke terms generated in the step (4). This technique tests various combinations of candidate smoke terms to determine which sets of smoke terms are most predictive in the curation set. The technique builds a list of smoke terms using greedy principles, always adding the term that provides the greatest improvement relative to the current solution. To avoid becoming satisfied with a local but nonglobal optimum, the heuristic allows for temporary moves in negative directions in hopes of ultimately reaching superior solutions; however, previous solutions are always remembered and retained in the event that a better solution is not found. Goldberg and Abrahams (2018) compare this technique with competing approaches and find that the Tabu search's ability to reduce the feature space from thousands of terms down to a more manageable quantity results in high-quality, interpretable solutions. We provide pseudocode for this approach in Fig. 3 below, and extended details may be found in Goldberg and Abrahams (2018).

(6) We evaluated the holdout set by using each of the text classification methods shown in Fig. 2. Each method was utilized to rank the reviews of holdout set from most likely to least likely to refer to a safety hazard. For smoke terms, we created a "smoke term score" for each review in the holdout set utilizing both the smoke terms and their prevalence scores from the CC score algorithm. Each time we found an instance of a smoke term in a text document, we incremented that review's smoke term score by that term's CC score. For example, suppose that the word "hurts" occurs twice in a text document, and its weight is 25,836. In this case, we would increment that text document's smoke term score by 2 occurrences $\times$ 25,836 weight $=$ 51,672. We applied this process for each term and for each text document. We then sorted all text documents in the holdout set from the highest to lowest by smoke term score, where the highest scores were more likely to refer to safety hazards. In addition to our machine-learned smoke terms, we also compared the performance of two popular sentiment analysis techniques, AFINN (Nielsen, 2011) and Harvard General Inquirer (Kelly & Stone, 1975) as a baseline, as sentiment analysis has been used for similar safety surveillance applications in the past (Yang, Yang, Jiang, & Zhang, 2012). For sentiment analyses, we used each sentiment method to determine the sentiment for each review on a numeric scale, where more positive numbers reflected more positive sentiment and more negative numbers reflected more negative sentiment. On the assumption that negative sentiment would be associated with safety hazards, we sorted the reviews from most negative sentiment to most positive sentiment, where the most negative scores were more likely to refer to safety hazards. Overall, better performing methods are those that return more safety

hazard reports in the top-ranking reviews. Finally, we compared the performance of each method to random chance.

4.3. Product-Level Risk Assessment

Operationalizing review-level risk assessments from text mining, we also assess risk at the product level. Classically, both in the literature (Li et al., 2018) and in major schema such as HACCP (Chatzopoulou et al., 2020; Wallace et al., 2014), product safety risk is assessed along two axes: likelihood (or frequency) and severity (or consequence). High likelihood implies increased chances that a consumer experiences an adverse reaction. For instance, some ingredients in food products may cause adverse reactions for a greater proportion of consumers than others. In addition, more widely circulated products expose larger numbers of consumers to that risk. High severity implies more dangerous consequences if a consumer experiences an adverse reaction. For example, some foodborne contaminants could necessitate hospital visits or even result in death, while other contaminants are more likely to cause mild discomfort.

We assess likelihood using two measures. First, to quantify the frequency of safety hazard reports for each product, we compute the proportion of reviews of the product that contain smoke terms, γ. A product in which 30% of reviews contain smoke terms is assessed as more likely to cause an adverse reaction than a product in which 20% of reviews contain smoke terms, for example. Second, as more widely circulated products increase the number of consumers that may be exposed to potential hazards, we also account for this consideration using each product's Amazon.com sales rank. We follow the approach suggested by Chevalier and Goolsbee (2003) and transform sales ranks using a Pareto distribution into estimated sales, σ. As a consolidated likelihood figure, we use the product of γ and σ.

To assess severity, we consider the smoke term scores for reviews of the product that contain smoke terms. As higher smoke term scores increase confidence in safety hazard designations, we use the magnitude of the smoke term scores as evidence for severity. Thus, for reviews that contain smoke terms, we compute a total smoke term score, where higher scores are more likely to refer to serious safety hazards. Thus, we assess severity as the average smoke term score across reviews containing smoke terms, φ.

Finally, we use as an overall risk score the product of the likelihood and severity estimates, or $\gamma \times \sigma \times \varphi$.

5. RESULTS

5.1. Text Mining Model

In Table I, we display the top 10 unigram, bigram, and trigram smoke terms generated by our smoke term techniques along with their corresponding relevance scores (weights) as determined by the CC scoring algorithm. Many of the smoke terms relate to symptoms of food poisoning, such as "diarrhea," "sick," "vomiting," etc. While these symptoms are likely uncomfortable, they may be excluded from many sentiment analysis dictionaries as they are more factual reports of events. Interestingly, some of the terms seem to invoke a narrative discussing the consumer's experience after becoming sick, such as smoke terms such as "hours," "later," or "night." This trend was especially true in bigrams and trigrams, which included phrases like "hours later," "i woke up," and "the next day." While none of these phrases specifically indicates a safety hazard, the use of this type of language in the context of these online posts is a very good predictor of safety hazard-related discussions.

To assess the quality of each technique's rankings, we first assessed our holdout set and compared the top 200-ranked reviews to the bottom 200-ranked reviews for each method. If a method is high-performing, then we expect it to retrieve many safety hazard reports in the top 200-ranked reviews and few safety hazard reports in the bottom 200-ranked reviews. We show each method's performance in the top and bottom 200-ranked reviews in Table II. Smoke term unigrams, bigrams, and trigrams were perfect or near-perfect in top-ranking reviews, detecting 200.0, 198.7, and 198.0 safety hazards, respectively (recall that results reflect an average of k-fold cross-validation, allowing for decimal values). By comparison, AFINN narrowly outperformed random chance, while Harvard GI was substantially worse than random chance, indicating that sentiment was a generally poor predictor of safety hazard reports. Smoke term unigrams performed the best in the bottom 200-ranked reviews, followed by smoke term bigrams and smoke term trigrams, respectively. Both AFINN and Harvard GI outperformed random chance in these bottom-ranked reviews, although the smoke term methods were superior.

Table I. Top Smoke Terms

Unigrams		Bigrams		Trigrams	
Term	Weight	Term	Weight	Term	Weight
Diarrhea	130,967	Hours later	59,120	Hours later I	38,071
Sick	96,150	And diarrhea	52,008	I woke up	37,134
Vomiting	95,267	Throwing up	48,334	Vomiting and diarrhea	36,138
Hours	92,035	Diarrhea and	47,442	The next day	34,627
Stomach	89,816	Got sick	46,361	To the bathroom	34,441
Nausea	76,344	Vomiting and	44,510	Hours after eating	33,279
Later	74,335	My stomach	43,165	Diarrhea and vomiting	31,292
Night	70,221	Food poisoning	42,299	An hour later	29,912
Fever	54,656	Had diarrhea	41,268	The next morning	27,342
Symptoms	52,704	Stomach cramps	39,707	I started feeling	26,084

Table II. Number of Safety-Hazard-Tagged Reviews and nDCG in Top 200-Ranked and Bottom 200-Ranked Reviews for Each Method

	Method	Safety Hazards	No Safety Hazards	nDCG
Smoke term list	*Unigrams*			
	Top 200	200.0	0.0	1.00
	Bottom 200	13.3	186.7	0.91
	Bigrams			
	Top 200	198.7	1.3	0.99
	Bottom 200	18.0	182.0	0.88
	Trigrams			
	Top 200	198.0	2.0	0.99
	Bottom 200	19.3	180.7	0.85
Sentiment analysis	*AFINN*			
	Top 200	89.3	110.7	0.53
	Bottom 200	56.3	143.7	0.73
	Harvard GI			
	Top 200	16.3	183.7	0.07
	Bottom 200	89.0	111.0	0.55
Random chance	*Random chance*			
	Top 200	89.2	110.8	0.45
	Bottom 200	89.2	110.8	0.55

We used a further metric, normalized discounted cumulative gain (nDCG) to assess the ranking quality of the top 200-ranked and bottom 200-ranked reviews for each method. As opposed to a simple count of the number of safety hazards versus no safety hazards found, this method utilizes a logarithmic smoothing function to weight the position of each record such that higher ranking records are emphasized (Järvelin & Kekäläinen, 2002).[3] The best possible ranking yields an nDCG of +1, while the worst possible ranking yields an nDCG of 0. As Table II illustrates, the smoke term methods also far outperformed sentiment analysis via nDCG in both the top 200-ranked and bottom 200-ranked reviews. In the top 200-ranked reviews, the Harvard GI method performed worse than random chance by this metric. Although all the smoke term methods

[3]Consider a ranking of N reviews. Discounted cumulative gain is defined as $DCG_N = \sum_{i=1}^{N}(\frac{rel_i}{\log_2(i+1)})$, where $rel_i = 1$ if review i is of the target classification ("safety hazard" for top-ranked reviews and "no safety hazard" for bottom-ranked reviews) and 0 otherwise. Idealized discounted cumulative gain is the discounted cumulative gain that an optimal ranking would yield, or $IDCG_N = \sum_{i=1}^{N}(\frac{1}{log_2(i+1)})$. Normalized discounted cumulative gain, or nDCG, is simply the ratio of the discounted cumulative gain at a particular ranking to the idealized discounted cumulative gain at that ranking. Thus, $nDCG_N = \frac{DCG_N}{IDCG_N}$.

Table III. Precision, Recall, and Lift for Each Method

Top *N*-Ranked Reviews	Precision/Recall/Lift				
	Smoke Term Unigrams	Smoke Term Bigrams	Smoke Term Trigrams	AFINN	Harvard GI
100	1.000/	1.000/	0.990/	0.080/	0.440/
	0.034/	0.034/	0.034/	0.003/	0.015/
	2.243	2.243	2.221	0.179	0.987
200	1.000/	0.994/	0.990/	0.080/	0.450/
	0.068/	0.068/	0.067/	0.005/	0.031/
	2.243	2.229	2.221	0.179	1.009
500	1.000/	0.998/	0.994/	0.378/	0.436/
	0.170/	0.170/	0.169/	0.064/	0.074/
	2.243	2.239	2.230	0.848	0.978
1,000	0.999/	0.997/	0.989/	0.508/	0.608/
	0.340/	0.339/	0.336/	0.173/	0.207/
	2.241	2.237	2.219	1.140	1.364
1,500	0.995/	0.997/	0.977/	0.518/	0.665/
	0.507/	0.509/	0.499/	0.264/	0.339/
	2.231	2.237	2.192	1.162	1.491

performed well, the smoke term unigrams offered the best performance overall[4].

For each method, we compared the portion of safety hazards in the top 200-ranked versus bottom 200-ranked reviews using a chi-squared (Wald) test. For each method, we found that there was a statistically significant difference between the portion of safety hazards in the top 200-ranked reviews and the portion in the bottom 200-ranked reviews at the 0.001 level. While the Harvard GI method's performance significantly differed, this was actually because the bottom 200-ranked reviews contained more safety hazard reports than the top 200-ranked reviews.

In addition, we utilize Kendall's Tau, a rank correlation coefficient, to examine the concordance between our methods' rankings and optimal rankings. The best possible ranking yields Kendall's Tau of $+1$, while the worst possible ranking yields Kendall's Tau of -1. Smoke term unigrams, bigrams, and trigrams achieved scores of $+0.85$, $+0.84$, and $+0.83$, respectively. The Harvard GI and AFINN sentiment methods also achieved positive correlations, but their scores of $+0.15$ and $+0.39$ were far weaker.

In Table III, we report precision, recall, and lift scores for each method at a variety of possible cutoffs. In each cell, we assume that the top *N*-ranked reviews by a given method were classified as safety hazards, and the remainder of the reviews were classified as no safety hazards. Precision refers to the proportion of these *N*-ranked reviews that were actually tagged as safety hazards; recall refers to the proportion of all safety hazard tags in the holdout set that were captured in these *N*-ranked reviews; and lift is the ratio of the number of safety hazards captured to the number that would be expected by random chance. We observe that the smoke term methods perform especially well at lower cutoffs, and performance declines slightly at higher cutoffs as most safety hazard reports have been detected. At these lower cutoffs, precision and lift are especially high while recall is low; as the cutoff increases, recall also increases, but precision and lift decline. Smoke term unigrams performed best at most cutoffs examined via precision, recall, and lift; however, managers can ultimately choose the method and cutoff that they feel are most consistent with their circumstances.

[4]This analysis utilized an unbalanced dataset (44.6% safety hazards and 55.4% no safety hazards), and the CC score algorithm is robust for such purposes. However, these proportions are so similar to balanced that we found only minor differences when experimenting with a balanced dataset. Unigram, bigram, and trigram methods detected 200.0, 199.0, and 198.3 safety hazards in the top 200-ranked reviews, achieving nDCG values of 1.00, 0.99, and 0.99, respectively. These scores were again superior to sentiment-based alternatives, where AFINN and Harvard GI detected 90.3 and 16.7 safety hazards, respectively, achieving nDCG values of 0.54 and 0.07.

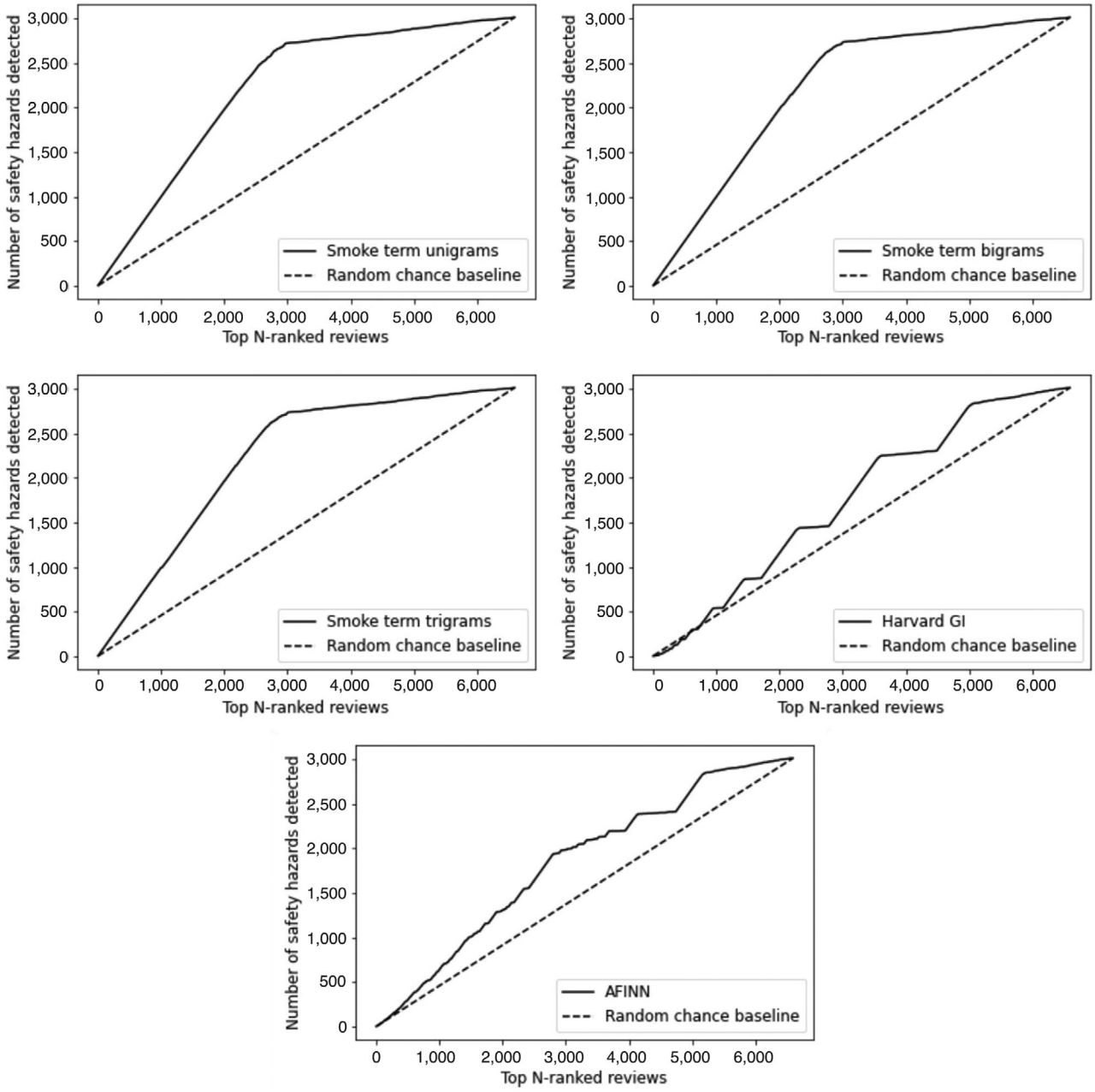

Fig 4. Graphical comparison of each method's performance.

Finally, in Fig. 4, we present lift charts that show the relationship between any arbitrary cutoff of the top *N*-ranked reviews and the number of safety hazards detected via each of the methods discussed. We find that smoke term unigrams, bigrams, and trigrams are all extremely high-performing, particularly at lower cutoffs. Each sentiment method slightly outperformed random chance.

5.2. High-Volume Data Validation

To validate the performance of our techniques in a high-volume setting, we examined a further 4,426,170 Amazon.com "verified purchase" grocery and canned food reviews (Ni et al., 2019). All the reviews utilized in our analysis were collected during the same time frame, but as our model was not

Table IV. Number of Safety Hazard-Tagged Reviews and nDCG in Top 200-Ranked and Bottom 200-Ranked Reviews for Each Method

	Method	Safety hazards	No safety hazards	nDCG
Smoke term list	*Unigrams*			
	Top 200	180	20	0.93
	Bottom 200	1	199	1.00
	Bigrams			
	Top 200	154	46	0.82
	Bottom 200	1	199	1.00
	Trigrams			
	Top 200	155	45	0.82
	Bottom 200	1	199	1.00
Sentiment analysis	*AFINN*			
	Top 200	52	148	0.30
	Bottom 200	1	199	1.00
	Harvard GI			
	Top 200	22	178	0.13
	Bottom 200	1	199	0.99
Random chance	*Random chance*			
	Top 200	2	198	0.01
	Bottom 200	2	198	0.99

trained upon this validation data, we can use it as further assessment of its efficacy. This data set also represents the use case for our technique: as user-generated content is too voluminous to read entirely, we utilize the methods discussed in this article to provide firms and/or regulators a shortlist of high-priority action items. Using each of the techniques discussed previously (unigrams, bigrams, trigrams, AFINN, and Harvard GI), we ranked the 4,426,170 reviews from most likely to least likely to refer to a safety hazard. We selected the top 200-ranked and bottom 200-ranked reviews via each method for further analysis. Due to some overlap in the rankings of different methods, 1,313 unique reviews remained after removing duplicated reviews that appeared in the top-ranked or bottom-ranked segments for multiple methods. Five undergraduate business students generated a total of 1,475 tags across these 1,313 reviews. The student taggers select the same tag 94.9% of the time, corresponding to a Cohen's κ score of 0.90, which Landis and Koch (1977) rate as "almost perfect agreement" and Fleiss et al. (2013) rate as "excellent agreement." Additionally, when comparing the student tags to a professional food scientist with a doctoral degree specializing in food safety as an external authority, we observed 95.0% agreement with a Cohen's κ score of 0.90, which Landis and Koch (1977) rate as "almost perfect agreement" and Fleiss et al. (2013) rate as "excellent agreement." These values were encouraging and similar to those obtained in our initial analysis, so we are convinced that the tagging was of high quality.

In Table IV, we present the results of our analysis. We assume in this analysis that the latent rate of safety hazards in the "grocery and canned food" category on Amazon.com is about 2.0%, as indicated by our first round of tagging. While each of unigram, bigram, and trigram methods identified mostly safety hazards in the top-ranking reviews, the sentiment analysis methods were far less effective. For each method, the portion of safety hazards in the top 200-ranked versus bottom 200-ranked reviews significantly differed using a chi-squared (Wald) test. This effect was again most dramatic for the smoke term methods, which identified mostly true safety hazards in top-ranking reviews.

5.3. Product-Level Analysis

To validate our techniques on a product level, we computed risk scores for each product in the high-volume setting of 4,426,170 Amazon.com "verified purchase" reviews. For each of 264,633 unique products, we derived an overall risk scored based on the smoke term analysis and the product's sales rank. As the unigram smoke terms were the top-performing, we used these smoke terms to perform our product-level analyses.

Table V. Affiliations and Experience of Expert Panel

Affiliation	Count	Average Years of Experience
Academic	9	21.0
Government	2	19.5
Industry	10	27.3
All	21	23.9

We selected the top 60-ranked products that our approach identified as the most likely to be hazardous for further analysis. For a baseline comparison, we also selected 20 random products with at least 50 reviews. A panel of 21 experts in the food safety domain provided assessments of each product's food safety risk. Each expert was assigned a random set of 20 products for analysis, including some hazard-flagged products and some baseline products. Experts were asked to rate the likelihood of a hazard, the severity of the hazard, and the overall hazard risk on a scale of 1 (very low), 2 (low), 3 (moderate), 4 (high), or 5 (very high). To perform this analysis, each expert was presented with the product's name, brand, Amazon.com sales rank, number of Amazon.com reviews, Amazon.com description, and five curated reviews. For hazard-flagged products, these five reviews were the top-ranking reviews as identified by the smoke term unigrams. For the random products, these five reviews were random 1-star or 2-star reviews (out of 5 stars possible). Each product's Amazon.com page was hyperlinked to provide the experts an opportunity to further investigate the product if they desired.

The experts spanned academic (food science and safety researchers), government (food safety enforcement), and industry (food safety consulting) backgrounds, and the average expert had over 20 years of experience in the domain. A detailed cross-tabulation is presented in Table V. The majority of experts (11) had doctoral degrees, and many experts had received industry training in food safety, such as HACCP (10), Safe Quality Food (SQF) (4), and ServSafe (2).

Each product was assessed by at least three experts; on a product level, we compute scores for likelihood, severity, and overall risk by averaging the experts' evaluations. We show a visual comparison of hazard-flagged versus baseline products in Fig. 5, where each product's mean risk assessment score is depicted.

Table VI. Smoke Term-Curated Reviews for Recalled Products

Protein Powder (First Recalled in 2016)			Nutrition Bar (First Recalled in 2017)		
Review	Date	Stars	Review	Date	Stars
… for the two days I drank it I became violently ill … couldn't go out in public and almost couldn't go to work because it made me so sick. Gas and extreme bloating …	02/26/2015	1	… Product looked funny and tasted like MOLD. I nearly threw up.	04/02/2015	1
[Food Poisoning Alert] … I was puking my brains out …	04/21/2015	3	Gave me and my wife instant diarrhea.	10/22/2015	1
… After a couple days of having these shakes for breakfast, my stomach hurt for DAYS on end. Almost 3 weeks …	11/17/2015	2	… just as I was about to bite into it I noticed that there was Human hair in it …	12/14/2015	1
… it made me bloated and my stomach cramped up.	12/03/2015	2	Moldy, gross. I threw up after taking a bite …	05/01/2016	1
… about a half hour after I finished my shake, I became extremely nauseous and ill …	12/05/2015	1	… very rancid even though Best By date was 11 months away …	11/04/2016	2

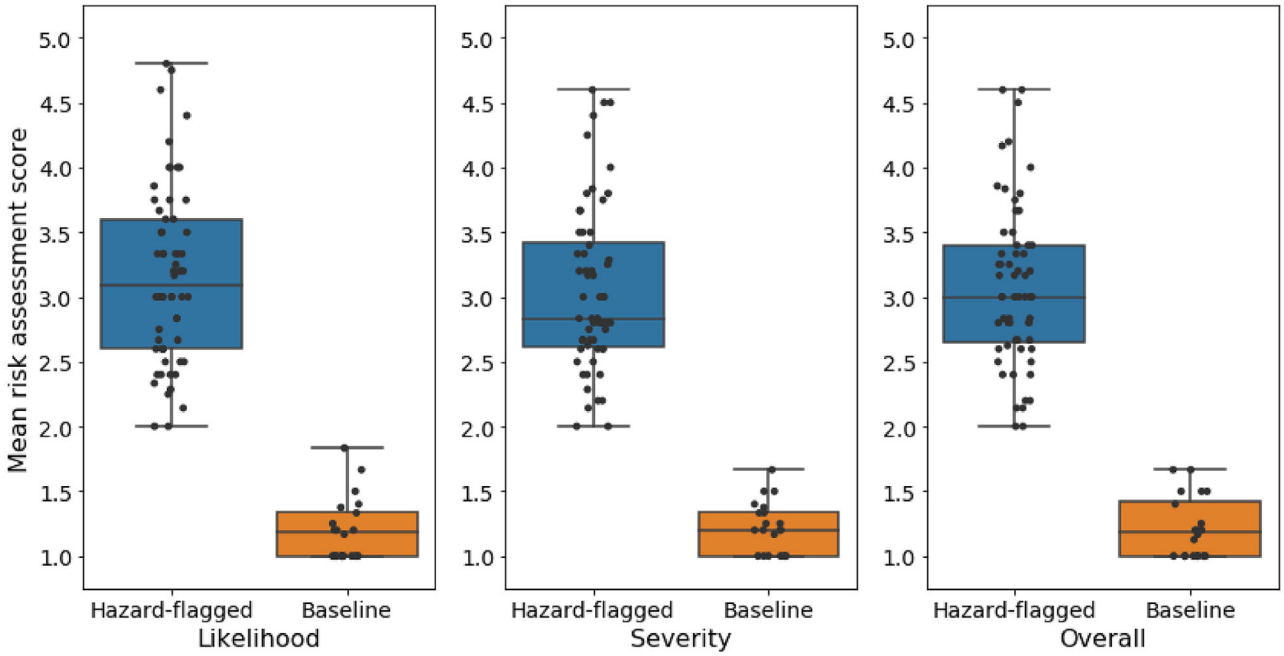

Fig 5. Graphical comparison of risk assessment scores for hazard-flagged products versus baseline products.

The experts rated the hazard-flagged products a mean (median) of 3.16 (3.08) on likelihood; 3.06 (2.83) on severity; and 3.09 (3.00) on overall risk. By contrast, the baseline products were assessed a mean (median) of 1.21 (1.19) on likelihood; 1.23 (1.20) on severity; and 1.22 (1.18) on overall risk. The baseline products had minimal spread, with standard deviations of 0.25, 0.20, and 0.24 on likelihood, severity, and overall risk, respectively. The hazard-flagged products had greater spread, with standard deviations of 0.68, 0.64, and 0.63 on likelihood, severity, and overall risk, respectively. The hazard-flagged products scored statistically significantly higher than the baseline products on each measure via right-tailed Mann–Whitney U tests at the 0.001 level, validating that the experts believed the hazard-flagged products to constitute elevated risks in comparison to the baseline products. Likelihood, severity, and overall risk ratings were highly positively correlated with one another. Likelihood and severity were correlated +0.88; likelihood and overall were correlated +0.93; and severity and overall were correlated +0.92.

Cross-referencing the hazard-flagged products' producers with FDA records, we were able to match 28 out of 60 producers. Some products, such as dietary supplements, do not require preapproval by the FDA, and thus, FDA records may not include their producers. Some other cases in which matches were not found could be attributable to subsidiaries using different names than their parent companies. Of the 28 matching products, we found that two products had been recalled. The first product, a protein powder, was recalled after being linked to a multistate outbreak of *Salmonella* (Gambino-Shirley et al., 2018). The second product, a nutrition bar, was recalled for failing to declare potential allergens on its packaging (KIND Dark Chocolate Nuts and Sea Salt Bar, 2017). Despite their eventual recalls, the protein powder averaged 3.7 stars in over 4,000 Amazon.com reviews, while the nutrition bar averaged 4.5 stars in over 11,000 Amazon.com reviews. Although few consumers reported that they were negatively affected by the quality issues leading to the recalls, the smoke terms were able to identify persistent safety concerns in each product that led to high-risk designations. In Table VI, we show some exemplar reviews for each product dated prior to the corresponding recalls as curated by the smoke terms. As some reviews were quite long, truncations are indicated by ellipses. In each case, the historical reviews suggest that further investigation of product quality practices could have been valuable for mitigating risk for future consumers. Moreover, given the great volume of reviews for each product, monitoring for signs of safety hazards proves difficult, and

the ability to more rapidly filter concerning reviews using smoke terms is invaluable.

As about 0.05% of cases of food poisoning are reported (Tack et al., 2019), it is unsurprising that the majority of the matching products were not recalled. However, given the rarity of recalls, the concordance between our results and FDA records is further support for the smoke terms' ability to identify real-world safety hazards and products subject to risk of recall. To this end, these results have been shared with FDA officials for further analysis.

6. DISCUSSION AND CONCLUSION

In this article, we propose a new FSMS for employing text mining to better understand online media in the food industry as firms and regulatory agencies seek to expediently detect safety hazards. We constructed a large labeled data set of reviews from both Amazon.com and IWasPoisoned.com. We then utilized supervised learning techniques to generate "smoke terms" that are predictive of mentions of safety hazards in these data sets. While the top smoke terms were nonemotive, they are extremely predictive of high-value online posts that an expert could quickly read and determine the best course for remediation, if necessary. We compared our analytics to traditional sentiment analysis methods, finding that the smoke terms offered fantastic performance, while more traditional methods lagged far behind. Thus, these smoke terms should serve as an actionable tool for better monitoring of online media for safety hazards in the food industry. On a product level, we present a novel approach for developing consolidated risk estimates, which align well with the judgments of food safety experts.

Our study is broadly consistent with other studies that have utilized similar smoke term methodologies in that using smoke terms specific to a product category tends to be more effective than sentiment analysis for detecting product safety hazards (Abrahams, Fan, et al., 2015; Abrahams, Jiao, et al., 2012; Adams et al., 2017; Goldberg & Abrahams, 2018; Mummalaneni et al., 2018). However, there is some variability in the effectiveness of smoke terms relative to sentiment analysis between product categories; in our work, we found that the gap in performance between smoke terms and sentiment analysis was rather stark, with smoke terms far outperforming sentiment analysis. In some other product categories, such as baby cribs, prior work has found that the relative performance of these techniques is more comparable (Mummalaneni et al., 2018). An analysis of why these techniques perform differently across product categories is an important question for future work to assess.

The smoke terms identified by our work are also relatively unique. There was no overlap between our smoke terms and the smoke terms identified in analyses of countertop appliances (Goldberg & Abrahams, 2018), baby cribs (Mummalaneni et al., 2018), or automobiles (Abrahams, Fan, et al., 2015; Abrahams, Jiao, et al., 2012). The smoke terms identified in our work tended to be rather specific to the product category, and as such terms such as "diarrhea" or "to the bathroom" would not intuitively generalize to those product categories. However, we did note some concordance between our smoke terms and those identified in a prior study on joint and muscle pain remedies (Adams et al., 2017), where the terms "diarrhea," "stomach," and "nausea" were shared. As both product categories may involve ingestion, it is intuitive that some smoke terms may be shared. As a result, it would also be interesting for future work to examine the applicability of our smoke terms to medication safety, which also includes ingestion.

It is also interesting to contrast our results with those of Tse et al. (2018), who analyzed social media text in response to a food recall scandal. As our research focused on hazard reports, we obtained predictive terms such as "diarrhea," "vomiting," "nausea," etc. Instead, Tse et al. (2018) observed that, after a recall, much of the social media posts concerned consumer sentiment on the company, including language on blame, lost trust, apologies, etc. Thus, it appears that there are key linguistic differences between the language of consumers who have interacted with hazardous food products and those who have heard about them in retrospect.

In practice, our proposed techniques would improve monitoring for firms and regulators to monitor product quality and preemptively mitigate potential risks. The great volume of online data generated each day is impossible for a single organization to manually process in its entirety, but organizations are prone to substantial risk if they do not make use of the data. Indeed, our study found that about 2.0% of the Amazon.com reviews that we examined mentioned safety hazards, and the rarity of these reviews makes them difficult to identify. Implementing our techniques in an enterprise setting would allow organizations to obtain a daily shortlist of potential action items. These items would be ranked from highest to lowest risk to allow managers to manage time

and resources efficiently between risk mitigation tasks. Our surveillance techniques would serve as an integral piece of an FSMS, allowing industry experts to focus on remediation. This approach would ensure that monitoring efforts are efficient while effectively managing risk factors.

Our work has implications for existing quantitative methods of quality control in operations management, such as Six Sigma (Zu, Fredendall, & Douglas, 2008). Many modern quality control schemata rely on using metrics to discern potential production problems quickly, and then implementing measures to remediate. This process is sometimes termed DMAIC, or Define-Measure-Analyze-Improve-Control. Our methods of safety surveillance could be integrated into the "measure" phase of this methodology, where continuous analysis of online media would provide insights as to potential production problems. Once a problem is identified, then the cause can be determined ("analyze"), the process can be rectified ("improve"), and product can guard against repeating the problem in the future ("control"). One potential means for integration is through statistical process control, in which production processes are continually monitored using statistical methods such as moving averages (Panagiotidou & Tagaras, 2010). In our context, a firm could use a moving average of smoke term scores as an indicator of production problems; a sudden spike in this score could, for instance, reflect a contamination that could be quickly contained. Implementing this approach is one promising avenue for future research.

In addition, we are interested in exploring several further extensions of this study. We used supervised learning methods in this study to generate smoke terms that were predictive of safety hazards. In a future study, it would be interesting to further categorize these safety hazards, possibly using manual content analysis or a further text mining approach such as topic modeling. For instance, perhaps, some safety hazards are more dangerous than others, and this information could further improve our ranking of reviews for practitioners to examine. A further future research opportunity is that, in addition to the online posts examined in this study, it would be valuable to apply the smoke terms to further contexts and assess their performance in monitoring other platforms, such as social media sites.

Beyond our application in the food industry, our research also provides a roadmap for applications of text mining to assess risk in other domains. One vital application would be to pharmaceuticals, where safety hazards posed by medications may also be evident in online media. Goldberg and Abrahams (2018) used text mining of online reviews to perform safety surveillance for over-the-counter medication, but further study could considerably augment this analysis. For instance, prescription medication could be an important addition to over-the-counter medication, and online patient forums may be a valuable supplement to online review data sets. Furthermore, validation of text mining tools by examining concordance with expert opinions, as we have introduced in this article, improves the credibility of results. Finally, there may be some necessary overlap between these domains concerning drug–food interactions (Ryu, Kim, & Lee, 2018), and applying smoke terms from both domains may aid in discovery.

Further domains may also benefit from advanced text mining methods. For instance, in occupational health and safety, text mining could be used to perform automated screening of accident reports (Goh & Ubeynarayana, 2017). Smoke terms could be used to categorize reports based upon severity and injury type. In this context, as opposed to conventional consumer safety, text-based product risk assessment could inform business-to-business procurement decisions. The expert validation that we have suggested could be applied to profile the levels of risk posed by various occupational activities. A further application may be to decision analysis, where text mining could be used to identify sources of risk in textual conversations. For example, Brahma et al. (2020) apply text mining in mortgage domain to mitigate risk in loan processing, and further applications of smoke terms could provide interpretable decision support for mitigating business risks. Another application may be to security and defense, where text mining has been used to crawl the Dark Web for evidence of terror chatter (Qin, Zhou, Reid, Lai, & Chen, 2007). Smoke term analyses could augment these tasks by providing a set of interpretable words and phrases associated with terror chatter. Particularly, in this domain, concordance with experts would be vital to calibrate a model whose determinations were reflective of expert risk perceptions. Finally, we can also envision utilizing these text mining approaches as an aspect of risk communication. While we have discussed firms' use of text mining to mitigate risk, there is also potential for product safety risks to be presented to prospective consumers in real time at the point-of-sale. Future work based on such an approach would be uniquely preventative rather than reactive and could improve upon conventional consumer

advisories, which occur in post and often struggle to reach all affected consumers (Hora, Bapuji, & Roth, 2011). With many possibilities, we hope that our work provides inspiration for future projects performing text-based risk assessments across domains.

ACKNOWLEDGMENTS

The authors are grateful to Patrick Quade, founder of IWasPoisoned.com, for generously providing access to a data set of food safety reports that enabled this study. The authors are grateful to Kimberle Badinelli, Hospitality Systems LLC; Jason Bolton, University of Maine; Scott Brooks, Tyson Foods Inc.; Martin Bucknavage, Pennsylvania State University; Klodian Dauti, Food Safety and Quality Experts Inc.; Yaohua Feng, Purdue University, Jeff Kronenberg, Food Safety Northwest LLC; Changqi Liu, San Diego State University; Londa Nwadike, Kansas State University and University of Missouri; Reza Ovissipour, Virginia Tech; Fred Reimers, Creative FoodSafe Solutions; Cesar Romero, US Air Force Public Health; Donna F. Schaffner, Rutgers University Food Innovation Center; Au Vo, Loyola Marymount University; and further anonymous experts for lending their domain knowledge in assessing the risks of food products for this study. The authors are grateful to H. Lester Schonberger, Virginia Tech; and further anonymous experts for lending their domain knowledge in a pilot version of the study. Finally, the authors are grateful to the special issue editors and anonymous reviewers, each of whom provided considerate, constructive, and invaluable feedback.

LITERATURE CITED

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

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems*, *55*(4), 871–882.

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

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

Behnke, K., & Janssen, M. (2020). Boundary conditions for traceability in food supply chains using blockchain technology. *International Journal of Information Management*, *52*, 101969.

Brahma, A., Goldberg, D. M., Zaman, N., & Aloiso, M. (2020). Automated mortgage origination delay detection from textual conversations. *Decision Support Systems*, *140*, 113433.

Chatzopoulou, S., Eriksson, N. L., & Eriksson, D. (2020). Improving risk assessment in the European food safety authority: Lessons from the European Medicines Agency. *Frontiers in Plant Science*, *11*, 349.

Chevalier, J., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics*, *1*(2), 203–222.

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

Cox Jr,, L. A., Popken, D. A., Sun, J., Liao, X. p., & Fang, L. X. (2020). Quantifying human health risks from Virginiamycin use in food animals in China. *Risk Analysis*, *40*, 1244–1257.

Delen, D., & Zolbanin, H. M. (2018). The analytics paradigm in business research. *Journal of Business Research*, *90*, 186–195.

Ducharme, J. (2019). You're not imagining it: Food recalls are getting more common. Here's why. Time.

Duggirala, H. J., Tonning, J. M., Smith, E., Bright, R. A., Baker, J. D., Ball, R., … Bouri, K. (2015). Use of data mining at the Food and Drug Administration. *Journal of the American Medical Informatics Association*, *23*(2), 428–434.

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

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. New York: John Wiley & Sons.

Folkerts, H., & Koehorst, H. (1998). Challenges in international food supply chains: Vertical co-ordination in the European agribusiness and food industries. *British Food Journal*, *100*(8), 385–388.

FAO. (2017). Food safety risk management: Evidence-informed policies and decisions, considering multiple factors. Food and Agriculture Organization of the United Nations.

FSIS Food Recalls. (2015).

Gambino-Shirley, K. J., Tesfai, A., Schwensohn, C. A., Burnett, C., Smith, L., Wagner, J. M., … Updike, D. (2018). Multistate outbreak of Salmonella Virchow infections linked to a powdered meal replacement product—United States, 2015–2016. *Clinical Infectious Diseases*, *67*(6), 890–896.

Garre, A., Boué, G., Fernández, P. S., Membré, J. M., & Egea, J. A. (2019). Evaluation of multicriteria decision analysis algorithms in food safety: A case study on emerging zoonoses prioritization. *Risk Analysis*, *40*, 336–351.

Goh, Y. M., & Ubeynarayana, C. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, *108*, 122–130.

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

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

Hsieh, W. T., Ku, T., Wu, C. M., & Chou, S. C. T. (2012). Social event radar: A bilingual context mining and sentiment analysis summarization system. *Proceedings of the ACL 2012 System Demonstrations*.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Kamilaris, A., Fonts, A., & Prenafeta-Boldύ, F. X. (2019). The rise of blockchain technology in agriculture and food supply chains. *Trends in Food Science and Technology*, *91*, 640–652.

Kate, K., Chaudhari, S., Prapanca, A., & Kalagnanam, J. (2014). FoodSIS: A text mining system to improve the state of food safety in Singapore. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kelly, E. F., & Stone, P. J. (1975). *Computer recognition of English word senses* (Vol. *13*). Amsterdam: North-Holland.

Keyvanpour, M. R., & Imani, M. B. (2013). Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms. *Intelligent Data Analysis*, *17*(3), 367–385.

KIND Dark Chocolate Nuts and Sea Salt Bar. (2017). Retrieved from https://www.accessdata.fda.gov/scripts/ires/?Product=158714

Kshetri, N. (2018). Blockchain's roles in meeting key supply chain management objectives. *International Journal of Information Management*, *39*, 80–89.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

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

LeBlanc, D. I., Villeneuve, S., Beni, L. H., Otten, A., Fazil, A., McKellar, R., & Delaquis, P. (2015). A national produce supply chain database for food safety risk analysis. *Journal of Food Engineering*, *147*, 24–38.

Leuschner, R. G., Robinson, T. P., Hugas, M., Cocconcelli, P. S., Richard-Forget, F., Klein, G., … Richardson, M. (2010). Qualified presumption of safety (QPS): A generic risk assessment approach for biological agents notified to the European Food Safety Authority (EFSA). *Trends in Food Science and Technology*, *21*(9), 425–435.

Li, J., Bao, C., & Wu, D. (2018). How to design rating schemes of risk matrices: A sequential updating approach. *Risk Analysis*, *38*(1), 99–117.

Mokhtari, A., & Van Doren, J. M. (2019). An agent-based model for pathogen persistence and cross-contamination dynamics in a food facility. *Risk Analysis*, *39*(5), 992–1021.

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

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

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

Njage, P. M. K., Henri, C., Leekitcharoenphon, P., Mistou, M. Y., Hendriksen, R. S., & Hald, T. (2019). Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. *Risk Analysis*, *39*(6), 1397–1413.

Nsoesie, E. O., Kluberg, S. A., & Brownstein, J. S. (2014). Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. *Preventative Medicine*, *67*, 264–269.

Outbreak of E. coli Infections Linked to Ground Beef. (2019). Centers for Disease Control and Prevention.

Panagiotidou, S., & Tagaras, G. (2010). Statistical process control and condition-based maintenance: A meaningful relationship through data sharing. *Production & Operations Management*, *19*(2), 156–171.

Qin, J., Zhou, Y., Reid, E., Lai, G., & Chen, H. (2007). Analyzing terror campaigns on the internet: Technical sophistication, content richness, and Web interactivity. *International Journal of Human-Computer Studies*, *65*(1), 71–84.

Ryu, J. Y., Kim, H. U., & Lee, S. Y. (2018). Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, *15*, E4304–E4311.

Scallan, E., Griffin, P. M., Angulo, F. J., Tauxe, R. V., & Hoekstra, R. M. (2011). Foodborne illness acquired in the United States—Unspecified agents. *Emerging Infectious Diseases*, *17*(1), 16–22.

Singh, A., Shukla, N., & Mishra, N. (2018). Social media data analytics to improve supply chain management in food industries. *Transportation Research Part E: Logistics and Transportation Review*, *114*, 398–415.

Song, C., Guo, C., Hunt, K., & Zhuang, J. (2020). An analysis of public opinions regarding take-away food safety: A 2015–2018 case study on Sina Weibo. *Foods*, *9*(4), 511.

Staniškis, J. K. (2012). Sustainable consumption and production: How to make it possible. *Clean Technologies and Environmental Policy*, *14*(6), 1015–1022.

Tack, D. M., Marder, E. P., Griffin, P. M., Cieslak, P. R., Dunn, J., Hurd, S., … Ryan, P. (2019). Preliminary incidence and trends of infections with pathogens transmitted commonly through food—Foodborne Diseases Active Surveillance Network, 10 US Sites, 2015–2018. *Morbidity and Mortality Weekly Report*, *68*(16), 369.

Tao, D., Yang, P., & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety*, *19*(2), 875–894.

Trienekens, J., & Zuurbier, P. (2008). Quality and safety standards in the food industry, developments and challenges. *International Journal of Production Economics*, *113*(1), 107–122.

Tse, Y. K., Loh, H., Ding, J., & Zhang, M. (2018). An investigation of social media data during a product recall scandal. *Enterprise Information Systems*, *12*(6), 733–751.

Van Der Vorst, J. G., Tromp, S. O., & Zee, D. J. V. (2009). Simulation modelling for food supply chain redesign; integrated decision making on product quality, sustainability and logistics. *International Journal of Production Research*, *47*(23), 6611–6631.

Wallace, C. A., Holyoak, L., Powell, S. C., & Dykes, F. C. (2014). HACCP – The difficulty with hazard analysis. *Food Control*, *35*(1), 233–240.

Yang, C. C., Yang, H., Jiang, L., & Zhang, M. (2012). Social media mining for drug safety signal detection. *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*.

Yao, L., & Parlar, M. (2019). Product recall timing optimization using dynamic programming. *International Journal of Production Economics*, *210*, 1–14.

Zaman, N., Goldberg, D. M., Abrahams, A. S., & Essig, R. A. (2020). Facebook hospital reviews: Automated service quality detection and relationships with patient satisfaction. *Decision Sciences*.

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