Received: 22 August 2019 | Revised: 26 December 2019 | Accepted: 13 January 2020

DOI: 10.1111/1541-4337.12540

COMPREHENSIVE REVIEWS IN FOOD SCIENCE AND FOOD SAFETY

Comprehensive REVIEWS in Food Science and Food Safety WILEY

Utilization of text mining as a big data analysis tool for food science and nutrition

Dandan Tao[1] | **Pengkun Yang**[2] | **Hao Feng**[1]

[1]Department of Food Science and Human Nutrition, College of Agricultural, Consumer and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois

[2]Department of Electrical Engineering, Princeton University, Princeton, New Jersey

Correspondence
Hao Feng, Department of Food Science and Human Nutrition, College of Agricultural, Consumer and Environmental Sciences, University of Illinois at Urbana-Champaign, 382F Agricultural Engineering Sciences Building, 1304 W. Pennsylvania Ave., Urbana, IL 61801.
Email: haofeng@illinois.edu

Funding information
Illinois Department of Agriculture, Grant/Award Number: IDOA SC-19-06

Abstract

Big data analysis has found applications in many industries due to its ability to turn huge amounts of data into insights for informed business and operational decisions. Advanced data mining techniques have been applied in many sectors of supply chains in the food industry. However, the previous work has mainly focused on the analysis of instrument-generated data such as those from hyperspectral imaging, spectroscopy, and biometric receptors. The importance of digital text data in the food and nutrition has only recently gained attention due to advancements in big data analytics. The purpose of this review is to provide an overview of the data sources, computational methods, and applications of text data in the food industry. Text mining techniques such as word-level analysis (e.g., frequency analysis), word association analysis (e.g., network analysis), and advanced techniques (e.g., text classification, text clustering, topic modeling, information retrieval, and sentiment analysis) will be discussed. Applications of text data analysis will be illustrated with respect to food safety and food fraud surveillance, dietary pattern characterization, consumer-opinion mining, new-product development, food knowledge discovery, food supply-chain management, and online food services. The goal is to provide insights for intelligent decision-making to improve food production, food safety, and human nutrition.

KEYWORDS

big data, information technology, semantic web, text mining

1 | INTRODUCTION

The big data era is driven by the explosion of data in nearly all sectors of our society, and it is a trendy topic affecting many aspects of our lives (Marvin, Janssen, Bouzembrak, Hendriksen, & Staats, 2017). The term "big" often refers to the data that are high in volume, velocity, variety, value, and veracity (5 Vs). As large amounts of data are generated in the era of digitalization, big data analysis has provided unprecedented opportunities in many areas such as chemical process industry, food industry, and pharmaceutical industry (Chiang, Lu, & Castillo, 2017). Data mining derived from the concept of "knowledge discovery from the data (KDD)" has been widely used in big data analysis (Han, Pei, & Kamber, 2011). Data mining is a collection of techniques in the process of knowledge discovery. As illustrated in Figure 1, three major stages of data mining are data preprocessing (to prepare targeted data to be mined), data pattern discovery (to identify patterns of the targeted data), and pattern evaluation and presentation (to identify and present the truly interesting patterns reparenting knowledge). Through data mining, patterns representing knowledge implicitly hidden in massive data can be automatically extracted for assisting human decision-making. A selective list of popular tools used for data mining can be found in Table 1.

Big data has been applied for production optimization and quality and safety assurance purposes through the food supply chain. One of the most fundamental challenges we are

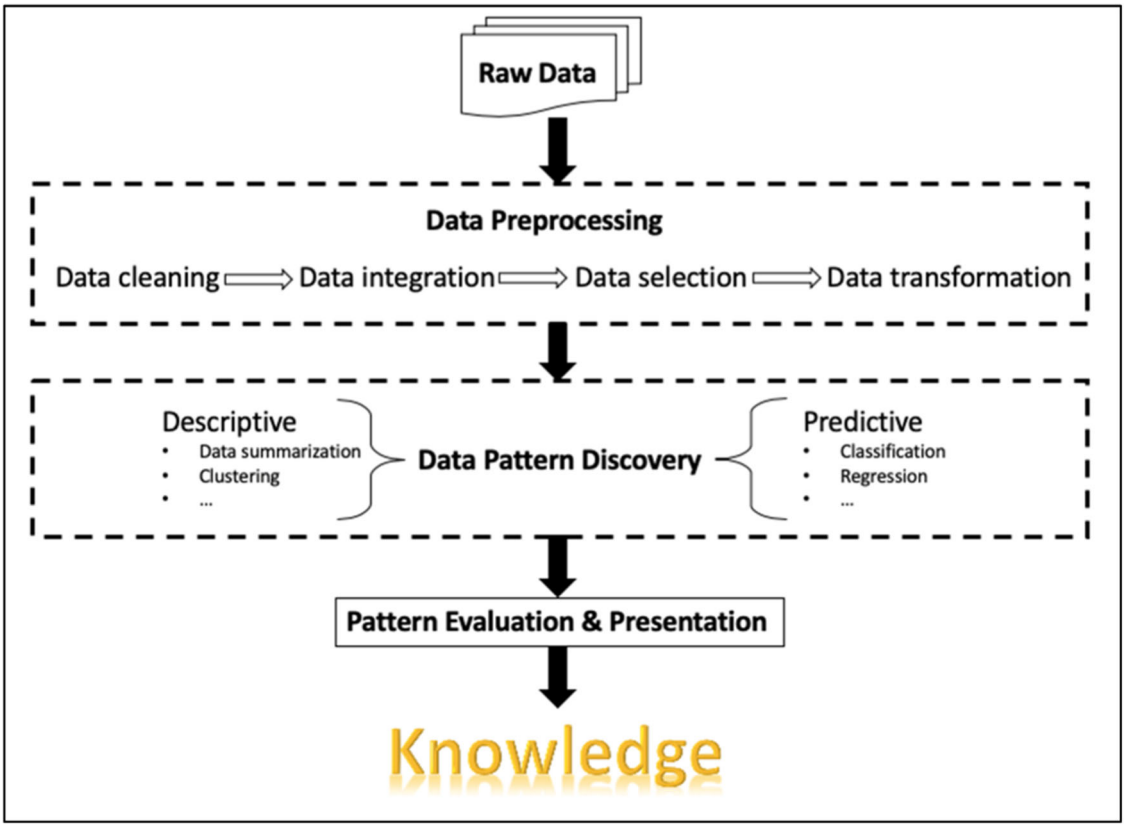

FIGURE 1 Data mining procedure in the process of knowledge discovery

TABLE 1 Popular software and packages used for data mining

Software/programming language	Library/package
Python	Pandas (Python)
R Software	Scikit-learn (Python)
SAS Enterprise Miner	NLTK (Python)
IBM SPSS modeler	NetworkX (Python)
Oracle Data Mining	Numpy (Python)
Orange Data Mining	tm (R)
RapidMinder	Sympy (Python)
Weka Data Mining	Scipy (Python)
Anaconda	nlp (R)
GNU Octave	wordcloud (R)
Gephi	apriori (R)
STATISTICA	topicmodels (R)
NVIVO	textir (R)
BigML	network (R)

Note. Many tools are not listed here; this list only focuses on data mining tools commonly used in text data analysis.

facing today is how to feed continuing growing global population with limited resources of land, water, and energy in the near future (Godfray et al., 2010). Ideas of "precision agriculture" and "smart farming" are proposed for improving agriculture production through monitoring, modeling, and optimized operations. The availability of huge amounts of data from multiple sources and sensors, as well as advancement of data storage and analyzing technologies are making big data analysis being rapidly adopted in agriculture sector (Kamilaris, Fonts, & Prenafeta-Boldó, 2019). For example, in food production, big data analysis has been used to make predictive insights about farm operations through models such as predictive yield model using Geographic Information System techniques (Al-Gaadi et al., 2016). Specifically, several vegetation indices (VIs) were extracted from massive satellite image data by software (data preprocessing stage); then VIs were fed to build a regression model to predict potato yield (data pattern discovery stage); the results were evaluated with actual yields (pattern evaluation stage). In the food process industry, quality is critical in influencing consumer acceptance of the final product (Du & Sun, 2006). To assure food quality through the supply chain, diverse sensors (e.g., hyperspectral imaging, spectroscopy, and biometric receptors) coupled with multivariate analysis methods have been used for classification and prediction purposes to evaluate food quality and authenticity (Jiménez-Carvelo, González-Casado, Bagur-González, & Cuadros-Rodríguez, 2019; Ropodi, Panagou, & Nychas, 2016).

Food is an essential component of our lives, cultures, and well-being (Abbar, Mejova, & Weber, 2015). The ways in which food is produced, prepared, and eaten is becoming more interactive and creative as more digital and network techniques are adopted by the public. The emergence of information technologies has allowed enormous

quantities of data to be collected and analyzed. As food is one of the most prevalent subjects in our lives, immeasurable amounts of food-related information are generated globally on a daily basis. Recently, the semantic web (e.g., Web sites, social media, online databases) has produced an increasing amount of digital data related to food production, food processing, and consumption. Digital platforms such as social media provide new possibilities for people to share their food consumption habits with others. Nowadays, it is not uncommon to see people in restaurants taking photographs of attractive dishes and immediately sending them on social media to share with others (Masson, Bubendorff, & Fraïssé, 2018). Scientists in many areas have become interested in identifying human food choices and consumption behaviors (Mouritsen, Edwards-Stuart, Ahn, & Ahnert, 2017).

Digital data are presented in a variety of forms, such as texts, images, and videos. Among them, text data play an essential role in our life, and have been studied most extensively (Paul & Dredze, 2017). Large amounts of text data are produced and consumed for communication purposes. Also, text data are generally rich in semantic content, containing people's knowledge, opinions, and preferences. They have, therefore, been used in a wide range of studies, in areas such as business intelligence, biomedical–literature mining, public health surveillance, agricultural management, and so on (Drury & Roche, 2019; Zhai & Massung, 2016). As more and more text information becomes accessible online, it has the potential to provide new insights, assist in decision-making, and improve product and service quality (Marvin et al., 2017). However, text data are usually extremely unstructured and difficult to analyze. New techniques have been developed to handle such data from diverse sources. These techniques are often referred to as text mining, a data mining division that focuses on discovering knowledge from text information (Zhai & Massung, 2016). From word-frequency analysis to advanced natural-language processing, text mining has created alternative methods of study and new insights in food-related topics. The previous work in data mining has mainly concentrated on instrument-generated data in the food industry (Chiang et al., 2017; Marvin et al., 2017; Ropodi et al., 2016; Waldner, 2017). To our knowledge, no effort has been made to summarize the research activities on text data or their utilization in the food science and nutrition domains.

In this compilation, an effort will be made to analyze how and to what extent text data play a role in food systems. First, the method used for data collection and the basic findings from the data will be described. Then the primary sources of text data in the digital era and the standard methods used in analyzing them will be discussed. The applications of text mining are divided into seven categories to highlight the usage of the data and their relations to food-related topics. Finally, some opportunities and challenges for future studies will be proposed.

2 | TEXT DATA SOURCES

Text information, that is, in the form of natural-language text (e.g., English text), can be found in all web pages, social media (e.g., tweets), news, scientific literature, public records, and many other types of documents (Zhai & Massung, 2016). Text information relating to food and nutrition research primarily belong to three classifications: database data, Internet data, and social media data (Figure 2). Database data include information from government and scientific databases. Internet data consist of news articles and Web sites created in public media or professional organizations. Social-media data are user-generated contents that can be published by anyone in social networks (e.g., Twitter), forums (e.g., Reddit), or review Web sites (e.g., Yelp). In this section, these three kinds of information sources will be discussed regarding how they can be used to study food science and nutrition-related subjects.

2.1 | Database data

Researchers have shown considerable interest in mining hidden knowledge from databases since the end of the last century (Chen, Han, & Yu, 1996; Han et al., 2011). Recently, a number of databases have been used to enhance our understanding of food-related topics, such as identifying food safety and fraud events, and the interplay between food and disease in complicated food systems (Jensen, Panagiotou, & Kouskoumvekaki, 2014; Karaa, Mannai, Dey, Ashour, & Olariu, 2016; Marvin et al., 2017; Yang, Swaminathan, Sharma, Ketkar, & Jason, 2011). Bouzembrak and Marvin (2016), for instance, used a Rapid Alert System for Food and Feed (RASFF) database to identify and monitor the hazards of food fraud and food safety. Another source of text information is scientific databases, from which food safety hazards can be identified (Lucas Luijckx, van de Brug, Leeman, van der Vossen, & Cnossen, 2016; Van de Brug, Luijckx, Cnossen, & Houben, 2014). Studies were conducted to analyze relationships among food, genes, and illness using information from scientific abstracts (Jensen et al., 2014; Karaa et al., 2016; Yang et al., 2011). In the food industry, companies own business-related databases including information of food production, processing, and consumer feedbacks. However, most of the industry-sourced databases are private, and thus are difficult to retrieve and analyze. Recently, partial information such as ingredient list and nutrition table of food products are shared and can be accessed through government databases, for example, U.S. Department of Agriculture (USDA, 2019). Databases from government agencies, international

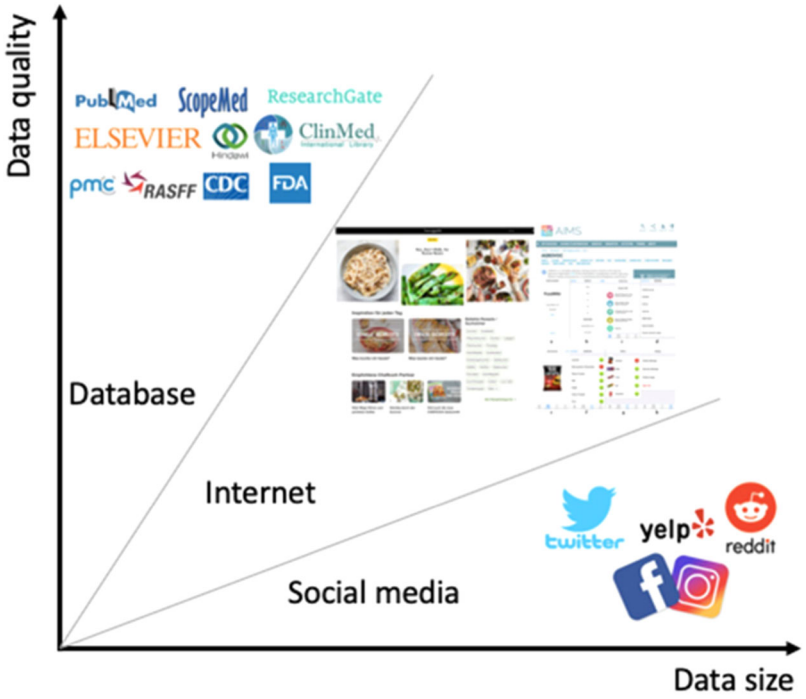

FIGURE 2 Text data sources used in food and nutrition applications

authorities, and scientific fields are usually high in data credibility.

2.2 | Internet data

The Internet is a rich source of food-related data (Marvin et al., 2017). Due to the wide implementation of digital techniques, news articles and professional reports related to food and nutrition are now easily accessible online. For example, food safety-related information scattered on the Internet is used for building database systems that can rank information based on their "relativity" to a food safety topic (Maeda, Kurita, and Ikeda, 2005; Kate, Chaudhari, Prapanca, & Kalagnanam, 2014; Chen, Huang, Nong, & Kwan, 2016). These surveillance systems are based on multisourced Internet data, such as mainstream news media, government Web sites, specialty blogs, and so on, and allow risk managers to get up-to-date information on food safety events (Steinberger, Pouliquen, & Van der Goot, 2013). In addition, text data on food composition, such as labels and recipes, are accessible on the Internet. They can be used to develop smart systems that can make predictive and adaptive decisions/suggestions (e.g., recipe completion and ingredient selection) based on data analytic algorithms (Ahnert, 2013; De Clercq, Stock, De Baets, & Waegeman, 2016). Published by official agencies or experts, these Internet-based data are usually high in relevancy and credibility.

2.3 | Social media data

In recent years, social media platforms like Twitter, Facebook, and Instagram have changed the way people interact and communicate with each other. These web-based microblogging platforms are generating real-time text data for the analysis of behavior, sentiments, and trends, and the surveillance of health matters (Ghosh & Guha, 2013). Consequently, there is a growing body of social media research focused on identifying the linguistic characteristics of the contents of food and human interactions. Scientists are using the text information to detect consumers' dietary patterns, adverse reactions, perceptions, preferences, and discussions on specific foods. The collection of dietary information includes maintaining diaries and regular surveys, but these are restricted in scope. However, the social media allow their users to update the world on the details of their daily lives, including their eating habits (Abbar et al., 2015). Web mining and social media analysis approaches are widely used for analyzing social media data. However, as the social media data are user-generated, the analysis of their text information still presents many challenges.

3 | TEXT DATA ANALYSIS METHODS

The fact that we are producing and consuming a lot of text data in communications indicates the importance of text data to our lives. In the days when we only needed to deal with small volumes of text data, manual processing was crucial and viable for improving productivity. With the fast increase in digital text information, manual processing, particularly for time-critical applications, is no longer feasible. Internet information comes in highly complex multivariate datasets and is

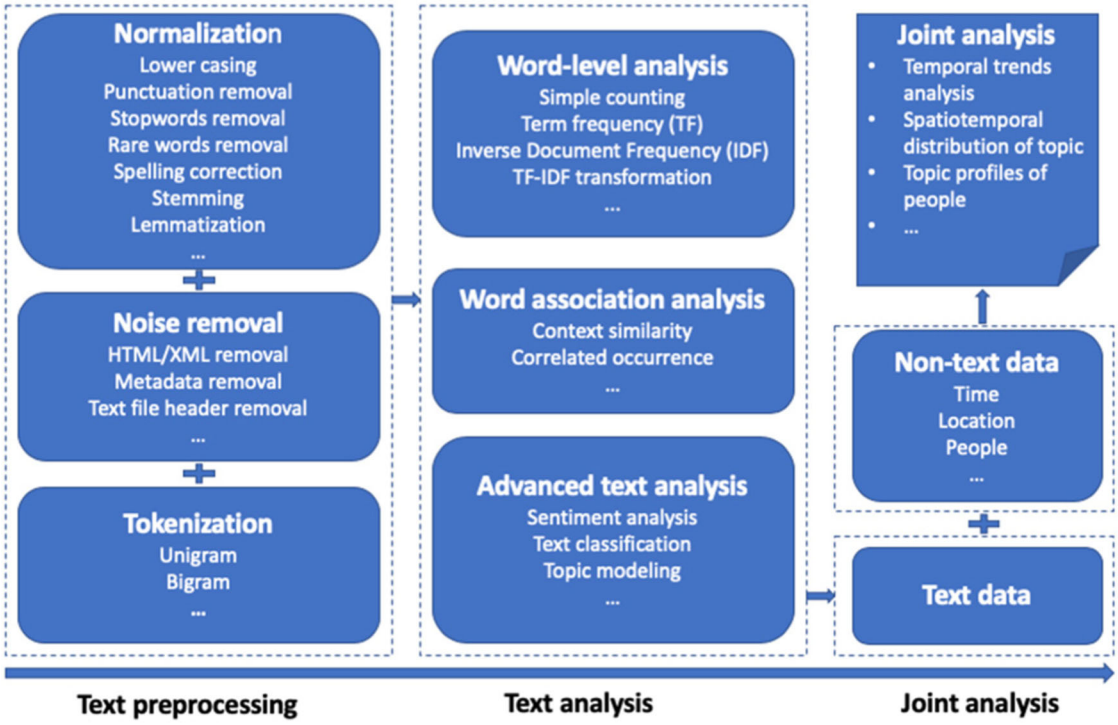

FIGURE 3 Schematic framework of text data analysis methods

difficult to inspect and present. A number of text analysis techniques, as referred to text mining, have therefore appeared to enable computers to transform large quantities of text information into useful insights. Text mining is a division of data mining that focuses on discovering knowledge from text information (Zhai & Massung, 2016). From data cleaning (basic text preprocessing methods), data reduction (word-level analysis), data association analysis (word association analysis) to advanced data mining (e.g., text classification, text clustering), the fundamental techniques used in text data analysis will be discussed in the following section. In addition, nontext metadata including demographics (times, places, incomes, etc.) can also be used to provide contextual information that is helpful in text analysis. To this end, standard methods used for the integration of text and nontext data will also be discussed. A schematic framework of these processes is shown in Figure 3.

3.1 | Basic text preprocessing methods

For many text mining tasks, data preprocessing is required. Normalization, noise removal, and tokenization are three general steps in preprocessing. *Normalization* refers to a series of tasks such as converting all letters to lower or upper case, converting numbers into words or removing them, removing punctuation, and so on. In particular, stop-word removal, stemming, and lemmatization are critical processes in text normalization. Words of high frequency, such as *I*, *the*, *of*, *my*, *it*, *to*, and *from*, which do not contain topical information, are called *stop words* (Zhai & Massung, 2016). They are usually removed in text analysis to improve the performance of the algorithm by reducing useless words in vector spaces. Words with different forms or derivationally related words with similar meanings are often reduced to a common base in text preprocessing (Manning, Raghavan, & Schütze, 2010). Stemming and lemmatization are two popular ways of removing such inflections. Stemming truncates the ends of words based on language-specific rules that often involve removing suffixes. In contrast, lemmatization utilizes vocabulary and morphological analyses to return a word's foundation (Manning et al., 2010). Both stemming and lemmatization, like stop-word removal, attempt to maintain the essential meaning behind the original text. Noise removal is another substitution method with more task-specific purposes, including the removal of HTML, XML, metadata, and headers from text files, the extraction of information from other formats, and so on. Tokenization generates an index over any text information that transforms single-text files into sparse vectors of term (or token) counts (Manning et al., 2010; Zhai & Massung, 2016). A successful tokenization system depends on a predefined dictionary of critical terms that depend on the particular issue (Zhai & Massung, 2016). A crucial token may be composed of a string of phrases, known as an *n*-gram of a series of *n* words, that captures the sequential relationship in the text information (Manning et al., 2010). A 2-gram (or bigram) is a two-word sequence of words like "I like," "like apple," or "apple juice," and a 3-gram (or trigram) is a three-word sequence of words like "I like apple," or "like apple juice" (Jurafsky &

Martin, 2008). Many promising applications have been discovered using bigrams and trigrams. Using longer grams offers more decision-making information but can also trigger data sparsity (Zhai & Massung, 2016). Although some task-specific processes may require manual coding to set the guidelines, most of the above functions can be achieved using current packages like "nlp" and "tm" in R programming (Hornik & Hornik, 2018), "nltk" in Python programming (Loper & Bird, 2002), and many other instruments like Weka (Hall et al., 2009), Stanford NLP (Manning et al., 2014), and MeTA (Massung, Geigle, & Zhai, 2016).

3.2 | Word-level analysis

Word-level analysis or frequency analysis is the most popular technique for handling text information (Zhai & Massung, 2016). This type of analysis is based on the frequency of occurrence of the tokens. In this approach, each document (sentence) has a concept to convey, and each different concept impacts the probability that tokens are used in the text (Hofmann, 2017). Word counting is the simplest method for analyzing frequency. However, frequency is often not based on simple counting, as different words need not necessarily provide the same amount of information. Weighted counting schemes have therefore been derived, such as Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how frequently a term occurs in a document, and IDF measures how important a word is in all documents (Zhai & Massung, 2016). A word's significance increases proportionally to the number of times it appears in a document, but this result is offset by its frequency in the corpus. The main weakness of word-level analysis is that the appearance of each token is assumed to depend only on the concept, and thus the word-sequence information is discarded. Nevertheless, owing to its simplicity, this strategy is very commonly used and has proven successful such as detecting influenza epidemics (Ginsberg et al., 2009). Also, critical-word sequences as in cases like negation can easily be adapted by incorporating the *n*-gram technique (Zhai & Massung, 2016).

3.3 | Word association analysis

In addition to word-level analysis, attention is given to define associations of words. Knowledge discovered from word association analysis is useful in many applications. For example, it can be used to suggest the variation of a query in text retrieval, or to construct a knowledge network using words as nodes and associations as edges. In particular, there is a growing interest in identifying associations of words related to foods, genes, and health (Ahn, Ahnert, Bagrow, & Barabási, 2011; Karaa et al., 2016). Network analysis is usually used to analyze graph-structured data for constructing knowledge systems. It is the science of understanding the connectivity of items in specific systems, and what the connections do. Networks of food-related words can be constructed from the association of word pairs. The last decade has seen applications of the network analysis of digital text data for improving our understanding of food science and nutrition. The examples of applying network analysis in food domain can be found in Section 4. For example, in Subsection 4.4, a flavor network was constructed based on online recipes and analyzed for providing ideas of creative cooking (Ahn, Ahnert, Bagrow, & Barabási, 2011; Simas, Ficek, Diaz-Guilera, Obrador, & Rodriguez, 2017). In Subsection 4.5, food disease and food nutrient networks were build based on literature text mining (Jensen et al., 2014; Karaa et al., 2016; Kim, Sung, Foo, Jin, & Kim, 2015; Yang et al., 2011).

3.4 | Advanced text analysis

Word-based analysis and word association mining can figure out the basic meaningful units in a language and how they are related to each other. However, advanced text analysis is required to determine the meaning of a sentence or a larger unit of a document. Advanced text analysis covers text classification, topic modeling, sentiment analysis, information retrieval (IR), and so on, including a variety of techniques under each of the topic (Miller, 1995; Zhai & Massung, 2016). In advanced text analysis tasks, machine learning (ML) techniques have been commonly used. ML is a paradigm in which a computer program learns how to make predictions or decisions from data based on two major procedures: training and testing (Ropodi et al., 2016). There are two kinds of ML, depending on whether the information is labeled or not: supervised learning and unsupervised learning. For example, text classification is a supervised ML task, whereas topic modeling is an unsupervised ML task. Nevertheless, the interpretation of the results requires the input from food scientists who are equipped with knowledge in statistics and data science. The commonly used text mining techniques in the food industry as well as ML algorithms used in the tasks are described below.

- Text classification: Text classification is a task to group text documents into specific classes according to a set of training records with predefined labels. The core of a text classification task is to train a classification model, which relates the features in the underlying record to one of the class labels (Aggarwal & Zhai, 2012). ML algorithms such as Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), and Neural Network are often used to build the classification models. Popular applications of text classification in our daily life includes spam detection, opinion mining, and information filtering. In the food industry, for example, text classification has been used to identify unreported cases of foodborne illnesses from social

media (Effland et al., 2018; Harris et al., 2014; Harrison et al., 2014; Sadilek et al., 2016).

- Text clustering: Text clustering is used for grouping objects (e.g., documents, paragraphs, sentences or terms) based on similarity between the objects (Aggarwal & Zhai, 2012). There are a variety of algorithms used for text clustering depending on how to calculate the similarity. Common methods include distance-based clustering methods, such as hierarchical clustering algorithms, partitioning algorithms, and the hybrid method using both hierarchical and partitional clustering algorithms (Aggarwal & Zhai, 2012). In the food industry, researchers have used text clustering to group food products based on consumer's review on specific attributes (Kim, Ha, Choi, & Moon, 2018; Lee, Ghimire, & Rho, 2013).
- Topic modeling: Topic modeling, as a text clustering division, detects potential topics in a document and often deals with information that does not have predefined labels (Zhai & Massung, 2016). A topic contains a cluster of words that frequently occur together. The main idea of topic modeling is to discover patterns of word use and how to connect documents that shared similar patterns. Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Probabilistic Latent Semantic Analysis (PLSA) are common ML algorithms used in topic modeling (Zhai & Massung, 2016). Topic modeling has met its applications for discovery of hidden semantic structures in a text body. In the food industry, topic modeling has been used to identify relevant public health topics such as obesity on Twitter (Ghosh & Guha, 2013).
- Sentiment analysis: Sentiment analysis (or opinion mining) is the computational study of people's affective states (e.g., opinions, emotions, attitudes) toward entities, issues, events, topics, and their attributes (Aggarwal & Zhai, 2012). For example, businesses always want to find consumer opinions about their products and services. Sentiment classification can be usually formulated as a supervised learning problem with three classes, positive, negative and neutral. ML algorithms such as NB and SVM are commonly used in sentiment analysis tasks. As opinion words and phrases are indicative for sentiment classification, unsupervised learning can also be used in sentiment analysis (Aggarwal & Zhai, 2012). Apart from classification of positive or negative sentiments, research has also been done on predicting the rating scores (e.g., 1 to 5 stars). In this case, the problem is formulated as regression problem. In the food industry, sentiment analysis has been used to know about the performances of food products and services from customer reviews (Gan, Ferns, Yu, & Jin, 2017; Hayashi, Hsieh, & Setiono, 2009).
- Information retrieval: IR is a task to find material (usually a text document) that satisfies information need (usually a keyword query) within large collections (usually a database; Manning et al., 2010). The core of an IR model is to assess the relevancy of a keyword query and a text document. Relevancy is determined by a similarity measure such as the cosine similarity that assumes that the queries and documents are represented as a vector of words (Drury & Roche, 2019). Effective IR models generally capture three heuristics, that is, TF weighting, IDF weighting, and document length normalization (Aggarwal and Zhai, 2012). Scientists have constructed a number of databases including text documents from various sources and built IR models for returning information most relevant to specific food safety topics (Maeda et al., 2005; Kate et al., 2014; Chen et al., 2016). The difference of algorithms in defining relevance has created a number of new ranking methods, such as ranking information based on published date or a score function (Chen et al., 2016), ranking with a formula adapted from Google ranking (Maeda et al., 2005), or ranking with a text classification method (Kate et al., 2014).

3.5 | Joint analysis of text and nontext data

In addition to text data analysis, it is beneficial to leverage nontext data in knowledge discovery from text data. Nontext can be used for providing background/context information (e.g., time, location, and people) for predictive analysis (Zhai & Massung, 2016). For example, the time and location information are useful "metadata" values of a text document. Given the context of time and location, it is possible to generate temporal or spatial trends of a particular topic discovered from text data (Zhai & Massung, 2016). In food-related studies, temporal analysis integrates frequency data with time and has been used in public health monitoring (Ginsberg et al., 2009). Spatial analysis shows the distribution of intensity across distinct places and has been used to characterize nutritional patterns (Abbar et al., 2015). On the other hand, text data can also help interpret patterns discovered from nontext data (Zhai & Massung, 2016).

4 | APPLICATION OF TEXT DATA IN FOOD-RELATED STUDIES

With food being one of the most common topics in our life, digital text analysis has been applied to a variety of food-related topics. In the following, the applications of digital text analysis related to food and nutrition will be discussed in seven categories of study: food safety and fraud surveillance, dietary pattern characterization, consumer opinion mining, new product development (NPD), food knowledge discovery, food supply chain management, and online food services. The publications discussed in this section are listed in Table S1 and summarized at the end of this section.

4.1 | Food safety and food fraud surveillance

Food safety plays a critical role in ensuring food security and sustainable food systems (Godfray et al., 2010; King et al., 2017). It is also a significant public health component. In the United States alone, an estimated 800 foodborne outbreaks are reported annually, accounting for about 15,000 illnesses, 800 hospitalizations, and 20 deaths according to the Centers for Disease Control and Prevention (CDC, 2018). Despite the development of several advanced surveillance and monitoring technologies, foodborne disease outbreaks remain to be a major threat to public health and the food industry. Food fraud is another problem that raises government concerns about food consumption under the scope of food safety (Spink & Moyer, 2011). Food supply chains and manufacturing methods become increasingly dynamic and complicated. Nowadays, it is not always appropriate to manage early detection and rapid response of food safety based on the results of science-based risk assessment (Maeda et al., 2005). Recently, the food industry has looked into data science and "big data" for insight into monitoring and responding in (near) real time to contamination threats as they occur (Greis and Nogueira, 2017). In particular, a variety of text information sources have been investigated to support risk detection and communication for food safety and fraud surveillance, including online database, the Internet, and social media (Marvin et al., 2017; Nsoesie, Kluberg, & Brownstein, 2014; Ordun et al., 2013; Tiozzo et al., 2019; Waldner, 2017).

For food risk and fraud surveillance, databases such as government databases and science databases, which are rich in text information, were leveraged. Thakur, Olafsson, Lee, and Hurburgh (2010) implemented a text classification algorithm DT to detect hidden trends in disease outbreaks and identify relationships between food kinds and outbreak places. They used text information from the CDC Outbreak Surveillance Data to identify vehicles and locations that were associated with specific etiologies. The resulting knowledge can help policymakers to inform successful food handling, preparation, and consumption practices interventions. Moore, Spink, and Lipp (2012) explored the scope, scale, and threat of food fraud problems from what was publicly reported in academic and media databases using keyword search. RASFF, created by European Commission, has enabled rapid exchange of food fraud-related information, so that governments can react quickly to protect consumers (RASFF, 2017). This established that RASFF system has assured food integrity through monitoring of both intentional and unintentional adulteration and fraud events (Esteki, Regueiro, & Simal-Gándara, 2019). Researchers have shown strong interest in identifying patterns of food fraud and predicting future fraud events by using data from the RASFF system. It has been used for predicting food fraud types, optimizing sample size for monitoring food safety risks, and identifying driving factors of food fraud events related to fruit and vegetables (Bouzembrak & Marvin, 2016; Bouzembrak, & Marvin, 2019; Bouzembrak, Camenzuli, Janssen, & Van der Fels-Klerx, 2018). In addition, scientific abstracts from MEDLINE/PubMed and FSTA databases were useful for designing an Emerging Risk Identification Support System (ERIS), which can identify unexpected hazards in the food chain (Lucas Luijckx et al., 2016; Van de Brug et al., 2014).

The Internet is another source of data for disease surveillance (Waldner, 2017). A variety of information systems around the globe has been developed to promote early warning of food safety and food fraud hazards through Internet data retrieval and text mining. For instance, a Japanese group built a database of documents on food safety hazard through keyword searching from Google web pages (Maeda et al., 2005). By visualizing the interactions of the documents, they also designed a Risk Path Finder system for people to recognize the emergence of a risk event that is hidden in a pile of documents. Singapore's National Environment Agency, in collaboration with IBM Research, created a Food Safety Information System (FoodSIS) to proactively monitor emerging food safety problems in Singapore using relevant food safety contents from the Internet (Kate et al., 2014). A database of food safety information was developed in 2016 based on food safety news from media and government Web sites to help efficiently assess food safety issues in China (Chen et al., 2016). The role of news media on food safety monitoring in China was further highlighted by focusing on the Chinese dairy sector (Zhu, Huang, & Manning, 2019). In Europe, Bouzembrak et al. (2018) developed a food fraud reporting system MeDISys-FF based on MeDISys, an infrastructure provided by the European Media Monitor that collects reports published worldwide in the media.

More recently, scientists are interested in employing new sources of digital text information to detect food safety and food fraud incidences. Open source media outlets such as Twitter and Rich Site Summary feeds were used to characterize the 2012 Salmonella event related to cantaloupes for predicting the number of sick, dead, and hospitalized (Ordun et al., 2013). Twitter and Yelp were employed to identify unreported foodborne illnesses in several local public health departments of the United States, and tested in cities such as Chicago, New York, and Las Vegas (Harris et al., 2014; Harrison et al., 2014; Effland et al., 2018; Sadilek, et al., 2016). Amazon reviews were analyzed with text classification methods for detecting issues of unsafe food products, with results validated by FDA food recalls (Maharana et al., 2019). Increasingly, the potential of employing social media data has gained the attention of governments for supporting efforts in public health surveillance. Besides detecting unreported foodborne diseases, social media has become a new channel for communicating food safety risks (Kuttschreuter et al., 2014; Meyer, Hamer, Terlau, Raithel, & Pongratz, 2015; Rutsaert

et al., 2013). Social media's popularity has made it a useful channel for customers to seek information on food safety, and for policymakers to communicate food safety crises to customers. Effective food safety management systems are critical to deal with the growing complexity and globalization of food supply chains. Digital data and solutions may provide new possibilities for predicting and controlling problems of food safety and food fraud to minimize economic losses (Fritsche, 2018).

4.2 | Dietary pattern characterization

In recent years, chronic diet-related illnesses triggered by unhealthy eating or imbalanced diet patterns have received growing attention worldwide. Studies were conducted to correlate nutritional profiles of individuals with their health to provide evidence and recommendations for efficient government health interventions. Previous studies on dietary pattern only gathered data from nutritional diaries or regular surveys, which are generally restricted in range and reach (Aiello, Schifanella, Quercia, & Del Prete, 2019). New information sources, such as social media now allow users to share their daily life with others, including diet and eating habits. Researchers are interested in employing these new sources of data for studying patterns of food consumption (Fried, Surdeanu, Kobourov, Hingle, & Bell, 2014). Among them, Twitter, Instagram, and Internet logs such as Google history are common sources of text information to characterize the nutritional pattern of people.

Real-time or archived text information from Twitter was used to study trends in health-related behaviors, consciousness, and monitoring. Ghosh and Guha (2013) identified relevant public health topics such as obesity on Twitter using topic modeling techniques and discovered the correlation between obesity-related tweets and prevalence of obesity rates among U.S. adults (with BMI $\geq$ 30; CDC, 2012). Using extra attribute information from tweets such as when and where the tweets are published could further extend the use of Twitter information in health studies. For instance, a technique for deriving dietary information in Twitter posts was suggested by Abbar et al. (2015). They noted a range of dietary information-derived correlations between Twitter and the incidence of obesity and diabetes in various U.S. counties. Based on tweets linked to distinct dining circumstances, Vidal, Ares, Machín, and Jaeger (2015) evaluated the shift in eating habits of people during different times of day. And Huang, Huang, and Nguyen (2019) used geotagged Twitter information to study the impacts of neighborhood features on dietary habits and the respective health results. García-León (2019) showed how Twitter hashtags reflected the food well-being of consumers in local food consumption.

Instagram is another platform used to study health-related subjects, including nutritional habits, owing to its growing popularity. Phan, Muralidhar, and Gatica-Perez (2019) used alcohol-related posts from Instagram to study the temporal, spatial, and contextual patterns of alcohol consumption in weekday nights. Sharma and De Choudhury (2015) focused on using Instagram to study ingestion practices and nutrition trends and identified how the wider Instagram community responds to low-calorie and high-calorie food. In specific, many studies have looked into the issue of food dessert, defined as regions with significant proportions of families with insufficient access to good food, using information from Instagram (Abbar et al., 2015; De Choudhury, Sharma, & Kiciman, 2016). These studies illustrated food decisions and dietary features in U.S. food deserts and highlighted the critical role of social media in defining the connection between eating patterns and their effect on people's health.

One's search queries related to food on the Internet could also be reflective of their dietary habits. Zhao et al. (2019) conducted a correlation analysis between Chinese dietary preferences from Baidu search data and diabetes risk from government statistics and found a geographical distribution pattern. Besides, online food recipes have been recognized as a good source of Internet data to study recipe production, consumption, and innovation (Rong, Liu, Huo, & Sun, 2019). West, White, and Horvitz (2013) and Wagner, Singer, and Strohmaier (2014) used text information from online recipes as a proxy to derive consumption and dietary patterns of individuals, indicating unbalanced food distribution and consumption that might be useful for avoidance of food-related health issues such as malnutrition and overnutrition. By tracking users' behavior of uploading recipes online, Trattner, Kusmierczyk, and Nørvåg (2019) revealed that one's social connections, in the form of friendship with other users, was predictive of what type of recipe the user will upload in the future. Asano and Biermann (2019) investigated dietary transition by analyzing a dataset of millions of recipes from the most popular German recipe Web site and found that a great number of people are shifting to plant-based diets with a 3.5% increase of vegan recipes submitted. This food transition pattern was confirmed by interviewing users showing extreme dietary change. In addition to these web-based data, database with billions of food purchase records has also been analyzed for identifying consumers' dietary patterns. Aiello et al. (2019) conducted a correlation analysis between food consumption indicated from digital records of grocery purchases and prevalence of obesity-related syndromes in London area. They found that increase of obesity rate was positively associated with calorie intake and negatively associated with nutrient diversity.

4.3 | Consumer opinion mining

The role of consumers in the process of NPD is becoming increasingly important in the food industry (Moskowitz & Saguy, 2013). Therefore, another line of research has explored consumers' opinions, primarily in text format, toward food and dining services for business intelligence. The traditional way to collect consumers' responses is through questionnaires or surveys, which is limited in scope and reach. Thus, the availability of consumer-generated Internet data is used as an option to extract useful food quality and preferences information. Recently, social media has been proven as a powerful tool for acquiring business insights and enabling intelligent product-development decisions. It has been employed to assist marketing in many food companies for years, and the most common application of social media is brand marketing. By comparing the quality and results of three products (Pizza Hut, Dominos Pizza, and Papa Johns Pizza), He, Zha, and Li (2013) provided competitive research on social media in the U.S. pizza sector. Alamsyah and Peranginangin (2015) analyzed brand communities of two giant companies (McDonald's and Burger King) in Indonesia to understand dynamic market behaviors of the regional fast food industry. Consumer reactions to product quality are essential for the food industry, in addition to brand management. Using content analysis, Ariyasriwatana and Quiroga (2016) classified "deliciousness" phrases from Yelp restaurant reviews, a popular social network that helps individuals make food decisions. By analyzing text information from online forums, Blackburn, Yilmaz, and Boyd (2018) and Masson et al. (2018) researched how individuals communicate about food on the Internet. More advanced text analysis, such as sentiment analysis, is often needed for a deeper understanding of customer preferences toward which food to eat or which restaurant to dine. For instance, Hayashi et al. (2009) used text mining and sentiment analysis to predict consumer preferences of fast food brands. Mostafa (2018) investigated people's sentiments toward halal through the study of 100,000 tweets from Twitter.com. And restaurant review sentiment analysis discovered the top three characteristics influencing restaurant customer preferences expressed as star scores as food, service, and context (Gan et al., 2017).

Similar to open-ended issues in sensory research, Internet text information is an alternative source for evaluating the sensory quality of food (Piqueras-Fiszman, 2015). The advantage of digital text data is that they are often free, spontaneous, and easy to get. In traditional sensory research, adjectives are commonly used to describe the sensory characteristics of food products. Lee et al. (2013) grouped tastes of 51 types of Korean cuisine and refreshments using 87 adjectives often used to describe Korean food tastes. They later extended the research by developing an automated text analysis method using ML for evaluation of food taste, smell, and characteristics from consumers' online reviews (Kim et al., 2018). McAuley and Leskovec (2013) modeled implicit taste preferences of customers and product characteristics from online food reviews with ML algorithms. The result indicated the possibility to construct a sensory-based food recommendation system that targets the correct products to the right individuals. Many recommender systems have been created to provide recommendations of food tailored to the individual taste of the user. Ge, Elahi, Fernaández-Tobías, Ricci, and Massimo (2015), for instance, built a tablet-based application that includes user-selected tags for rated food items to predict their prospective likings to new products. With an ever-increasing amount of food-related data being digitally recorded, it is expected that our perception, consumption, and choices of food will be significantly influenced (Mouritsen et al., 2017).

4.4 | New product development

Digital food composition and preparation data are used to assist research & development (R&D) procedures in the food sector for creating better formulations and speeding up product development process (Chiang et al., 2017). One of the most significant examples of text mining applications in the culinary sector is the discovery of flavor pairing patterns from millions of recipes. Ahn et al. (2011) carried out a network study on regional recipes and built a "flavor network." They discovered from the network that Western cuisines prefer to combine ingredients that share flavor compounds, while East Asian cuisines, particularly India, are prone to avoid it. Subsequently, a food bridging hypothesis was suggested to extend the study. The theory demonstrated that although two components do not share potent flavor compounds, they may become affine through a chain of pair affinities (Simas et al., 2017). On the other side, computational gastronomy's emergence and growth can add to novel combinations of ingredients and new product formulations (Ahnert, 2013; De Clercq et al., 2016). Accordingly, online recipe mining has inspired development of recommender systems with algorithms designed for giving users healthier suggestions (Trattner & Elsweiler, 2019). For example, Pinel, Varshney, and Bhattacharjya (2015) created a smart system to produce new recipes computationally, which later became known as IBM Chef Watson (Varshney et al., 2019). Chen et al. (2019) proposed a recommender system NutRec to provide suggestions on healthy recipes based on quantities of ingredients and their interactions. Nyati, Rawat, Gupta, Aggrawal, and Arora (2019) have shown a system for recipe recommendation based on the formation of ingredient–ingredient network and recipe–ingredient network. Without the implementation of computational approaches on these digital text data, regional patterns of culinary formulations cannot be easily discovered, nor the possibility of having intelligent systems giving us advice on how to prepare tasty meals. However, researchers have highlighted the importance

of the skill of the chef to ensure the palatability of the final dishes whose recipes were creatively generated by computers (Mouritsen et al., 2017; Spence, Wang, & Youssef, 2017).

R&D has always played a vital role in the food industry. The significance of integrating customer reaction to the phase of NPD has been emphasized (Moskowitz & Saguy, 2013). Instead of concentrating solely on sensory evaluations, it is suggested that future food businesses consider incorporating more feedbacks from the outsider, the consumer, into their NPD process. We have seen a growing amount of research now leveraging online digital data for marketing and consumer management. Christensen, Nørskov, Frederiksen, and Scholderer (2017) demonstrated the possibility of identifying new product ideas in online communities through text mining and ML. However, researches utilizing consumer data in the NPD process are rarely reported in the food domain. Carr et al. (2015) demonstrated a study employing social media data for identifying consumers' discussions around aroma, "coffee freshness." They found that social media not only generated product and category-related insights, but that those insights are reliable and in line with those derived from traditional research methods. The authors proposed that social media data should not only be used for marketing purposes, it could also be included in the process of NPD. Bashir, Papamichail, and Malik (2017), however, pointed out that adopting external ideas from social media for product development is not common in multinational companies. The value of using social media data for future innovation has been discussed recently (Bhimani, Mention, & Barlatier, 2019; Muninger, Hammedi, & Mahr, 2019). Muninger et al. (2019) highlighted challenges in applying social media in innovation, such as complexity of social media use and involvement of people from multiple groups to acquire and diffuse knowledge from social media.

4.5 | Food knowledge discovery

Understanding of the chemical and biological effects of food compositions on human health is critical to the nexus of food, nutrition, and health. Multiple sources of data are used for investigating food composition, including information from nutrition literature, product labels, food composition databases, and food regulations (Greenfield & Southgate, 2003). Recently, new computational methods such as network analysis and text mining have been implemented to manage knowledge of chemical components of foods of importance to human health. Product labels include valuable information about food compositions and health-related claims, which are primarily presented in text format. do Nascimento, Fiates, dos Anjos, and Teixeira (2013) analyzed ingredient lists of gluten-free food and food with gluten by mining ingredient list of commercial products. They found that the diversity of ingredients in gluten-free products is significantly lower. Roman, Belorio, and Gomez (2019) also conducted text analysis of ingredient list and nutrition facts of gluten-free products, revealing that commercial breads tend to use a combination of several starchy sources instead of a single ingredient for quality optimization. Fardet, Lakhssassi, and Briffaz (2018) categorized commercial food products based on the ingredient list, number of additives, texture, water activity, and shelf-life data.

Scientific literature is a significant source of information linked to food and nutrition sciences. Researchers have worked extensively to discover bioactive compounds in food and their relationships to varieties of diseases. The knowledge of how food elements impact human health, however, is restricted to the complex network. Recently, relationships are being built among food, gene, and diet-induced illnesses using advanced computational methods such as literature mining and network analysis (Jensen et al., 2014; Karaa et al., 2016; Yang et al., 2011). Also, Kim et al. (2015) constructed a food–food network with a node representing a food, and an edge between two nodes representing similarities of nutritional contents between the two foods. Jensen et al. (2014) created a Nutrichem database that connects plant-based foods with their small-molecule components and phenotypes of human disease. Anbarkhan, Stanier, and Sharp (2018) used text mining to obtain connections between obesity and herbal plants in the Arabic region. Rakhi, Tuwani, Mukherjee, and Bagler (2018) recognized the health benefits of culinary herbs and spices through literature mining. Chaix, Deléger, Bossy, and Nédellec (2019) applied text mining to a big collection of PubMed scientific paper abstracts to identify ecological diversity and the origin of microbial presence in food.

Ontology, defined as formal naming of a set of concepts within a domain, has been widely used for knowledge discovery in the age of Semantic Web (Eftimov, Ispirova, Potočnik, Ogrinc, & Seljak, 2019). As we move quickly toward the Internet of Things (IoT) paradigm, advancing food ontology would provide effective communications of food, ingredients, and health outcomes from a semantic view (Boulos, Yassine, Shirmohammadi, Namahoot, & Brückner, 2015). A few examples of food ontologies designed for various purposes are FoodWiki, AGROVOC, Open Food Facts, Food Product Ontology, and Foodon (Boulos et al., 2015; Dooley et al., 2018). For instance, FoodWiki is a system designed for customers to quickly examine the free text written on packaged food products for inferring their side effects (Çelik, 2015; Ertuğrul, 2016). As the advocacy of clean-label movements, individuals are more cautious about packaged foods with additives with potential health hazards nowadays. Applications like FoodWiki could assist customers to make wiser and healthier purchases. Foodon, another example of food ontology, has been created to increase global food traceability, quality control, and risk management (Dooley et al., 2018). ISO-FOOD ontology was created for sharing and organizing stable isotope

data across food science (Eftimov et al., 2019). ONE ontology was developed for enhancing reporting and communication of nutritional epidemiologic studies and data (Yang et al., 2019). And an ontology-based recipe repository that describes cooking terms and activities was developed for facilitating the sharing and searching of recipe data (Öztürk, & Özacar, 2019).

4.6 | Food supply chain management

The advances in the IoT and big data technologies have contributed to transform today's manufacturing paradigm to smart manufacturing (Tao, Qi, Liu, & Kusiak, 2018). With the ability of data collection from multiple stages in food supply chains, the IoT has made it possible for creating a more transparent, sustainable, and efficient food manufacturing (Astill et al., 2019). Blockchain, a technology that enables the sharing of encrypted records or digital events among collaborating parties, has recently been applied to agriculture and food retail supply chain for increasing food traceability (Astill et al., 2019; Kamilaris, Fonts, & Prenafeta-Boldó, 2019). The combined usage of the IoT and Blockchain technology would have the potential to improve supply chain management (Rejeb, Keogh, & Treiblmaier, 2019). For instance, a significant issue in food supply chain is food waste (Pearson & Perera, 2018). It was estimated that one third of the foods for human consumption is either lost or wasted throughout the food supply chain, from farmer to processing, transportation, retailing, and consumption (Ishangulyyev, Kim, & Lee, 2019). The IoT and Blockchain technologies have enabled more efficient information sharing and communication through the supply chain, which would facilitate the detection and prevention of food safety issues, and reducing the food waste (Minnens, Lucas Luijckx, & Verbeke, 2019). However, the majority of data leveraged in supply chain optimization are from quality sensing technologies such as RFID, bar code identifier, quick response code scanner, infrared sensors, video camera, and so on (Kodan, Parmar, & Pathania, 2019). The utilization of text data analytics for assisting food supply chain management is rarely reported. The application of text information from the semantic web in assisting agricultural decision-makings was emphasized in a recent survey (Drury & Roche, 2019). Also, the rapid development of business intelligence from consumer opinions might be useful for optimizing food production. For example, a case study of text mining on Twitter posts discovered main concerns related to beef products, which could be used for developing a consumer-centric beef supply chain that evolves with consumer needs (Singh, Shukla, & Mishra, 2018). In addition, digital technologies hold potential of increasing efficiencies within food retail supply chains, and transforming the food systems to become more sustainable (El Bilali & Allahyari, 2018).

4.7 | Online food services

With the rise of digital technology, online food delivery services are booming worldwide in recent years (Correa et al., 2019; Xu, & Huang, 2019). In particular, the development of food ordering/delivery mobile applications (e.g., UberEats, GrubHub, Dianping, Meituan, Ele.me) has changed the way in which customers interact with food services (Kapoor and Vij, 2018; Xu & Huang, 2019). The data generated from food delivery apps can be used for optimizing delivery route for reducing the time consumption of the consumers, identifying individual's ordering behavior online, and improving food services. For example, Jia (2018) examined restaurant customers' ratings and reviews, and identified high-frequency words, major topics, and subtopics using text mining. It has proven that digital text data analytics is a cost-effect approach for restaurants to gain quality improvement ideas from customers. In particular, topic modeling, sentiment analysis, and network analysis are popular text analysis methods used in mining customer comments (Ibrahim & Wang, 2019).

4.8 | Summary

Based on the discussion in this section, a core collection of 57 publications was further summarized by their publication time, publication type, data source used, and data analysis methods used (as shown in Table S1). These papers were all technical and research articles published from 2010 to 2019, including 45 journal papers and 12 conference papers. No review article or short communication article was included. The text data sources, and text analysis methods used in each paper were identified in the table.

We summarized the usage of different data sources in various applications in Figure 4. The total number of papers using a specific source of data (e.g., social media) was divided by the purposes/topics of study in each paper. Among all sources of data included in this work, social media has been the primary source of data being analyzed for identifying food integrity issues, consumers' opinions, and the dietary pattern of a population. The large volume of social media would provide unprecedented opportunities in future food innovation and production (Kosior, 2019). The Internet data such as online recipes are useful for investigating users' food preferences, designing food formulation, and refining food recommendation system. On the other hand, news media and government Web sites can be used for constructing systems for early detection of food safety/fraud and fraud hazards. Although database data have met their main application in discovering food and nutrition knowledge, they have also shown potential in preventing food safety or fraud issues. In addition, digital text data analytics can also be used for acquiring new insights of optimizing food operations, and minimizing food lost in supply chain.

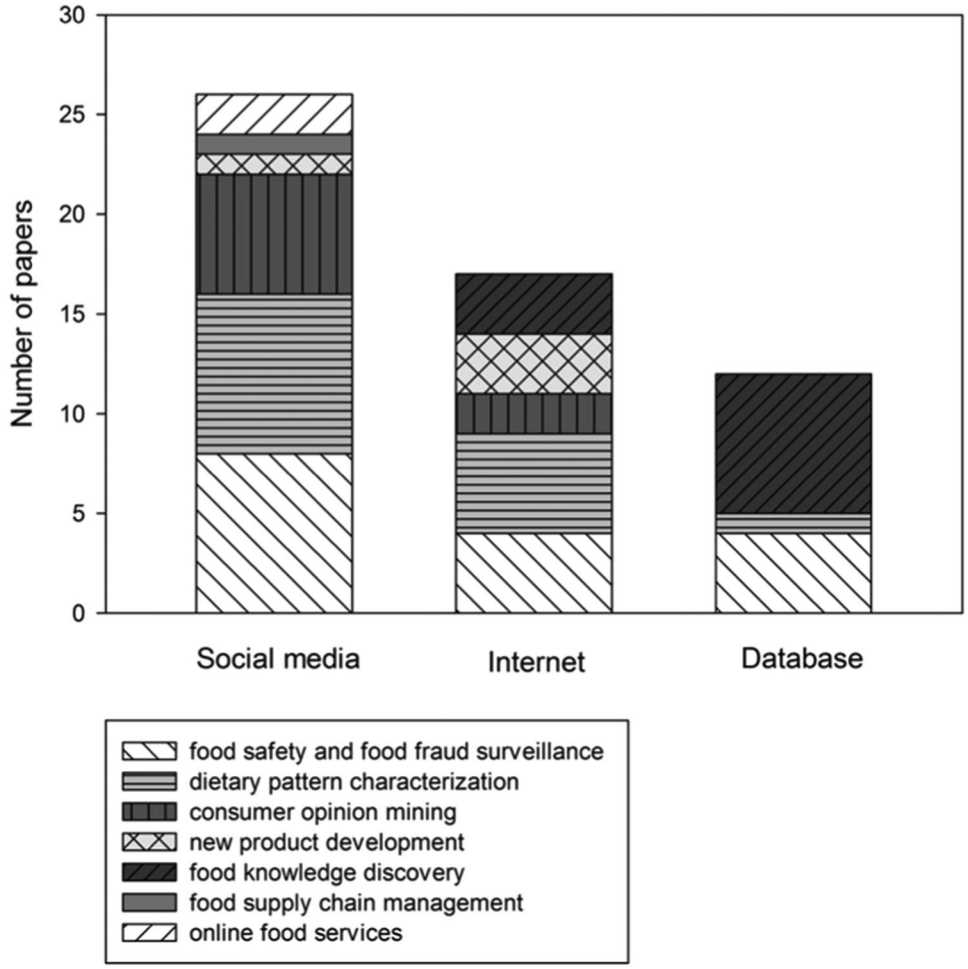

FIGURE 4 Distribution of text data sources in different fields of studies

The methods used in the fields of studies are shown in Figure 5. The total number of papers using a specific data analysis method (e.g., text classification) was also divided by the purposes/topics of study in each paper. A variety of text analysis methods have been applied to food-related studies. Word-level analysis is the most popular method followed by joint analysis of text and nontext data, with the largest proportion focusing on dietary pattern characterization. Word association analysis (e.g., network analysis) has been mostly applied to discover knowledge from massive documents. We have also seen the adoption of advanced text analysis methods such as text classification, IR, topic modelling, sentiment analysis, and text clustering. They are mainly used for detecting and prevention of food safety risks. Sentiment analysis has met its applications in helping understanding consumer preferences toward food product or food brand.

5 | OPPORTUNITIES

Big data has found extensive use in solving complicated real-world issues with strong applications in fields outside of computer science, such as chemistry, biomedicine, pharmacy, and agriculture (Chiang et al., 2017; Drury & Roche, 2019). With increasing number of food-related data produced on the semantic web, it is natural to think that the leverage of these data can benefit the food industry widely. This review includes, in specific, the use of text information in food-related research subjects. We have identified a number of applications of big data in the food industry. For example, food companies are innovating their NPD process (e.g., new flavor combinations) through analysis of multisourced data from ingredient lists, sensory results, and consumer preferences. Public health departments are developing new approaches for identifying people's consumption patterns for improving their surveillance of chronic diseases and foodborne outbreaks. Consumers are making food purchase or dining decisions based on recommendation systems that can provide suggestions based on their eating habits. We can foresee that text data analysis will meet its wider applications in the near future.

For example, we have seen the fusion of different sources of data helping to identify food safety and fraud hazards and characterize the consumption patterns of people in connection

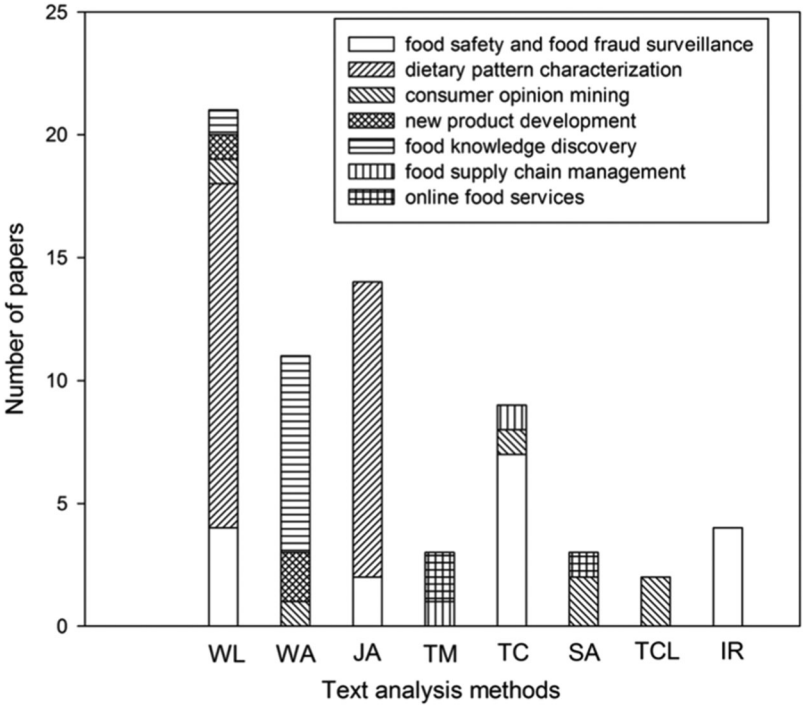

FIGURE 5 Distribution of text data analysis methods in different fields of studies (WL, word-level analysis; WA, word association analysis; JA, joint analysis of text and nontext data; TM, topic modeling; TC, text classification; SA, sentiment analysis; TCL, text clustering; IR, information retrieval)

with health such as obesity rate. As food has always been of great importance to public health, the potential of text data analytics in public health domain such as detection of food-borne illnesses and discovery of healthy and unhealthy dietary patterns have been shown in a variety of studies (Abbar et al., 2015; Harris et al., 2014; Harrison et al., 2014; Huang et al., 2019; Sadilek et al., 2016; Vidal et al., 2015). As a result, public health departments, in the future, may be able to identify safety and health hazards earlier to improve the performance of food administration (Duggirala et al., 2015).

Business intelligence is another area for utilization of digital text data. Web-based text information is used to help companies make more informed choices on brand management, customer preference, and NPD. Furthermore, target marketing and recommendation systems would allow distributors to provide the individuals with the right products based on their prior buying habits or preferential profiles. Digital text analytics can be viewed as a driver for the competitive benefits of food distributors by gathering, storing, and analyzing enormous amounts of customer information (Galletti & Papadimitriou, 2013). However, maintaining the balance between business intelligence and health outcome is a new challenge in the food industry that should not be ignored (Montgomery, Chester, Nixon, Levy, & Dorfman, 2019). In the meantime, these studies may also benefit customers. For example, a variety of information systems and smartphone applications have been created to help customers buy, prepare and manage their diets smarter and more personalized (Ahn et al., 2011; De Clercq et al., 2016; Ertuğrul, 2016).

6 | CHALLENGES

Unlike structured information gathered from instrumental detectors, text information collected from web pages, blogs, and social media are unstructured and hard to analyze. Although social media is the most promising source of data with extensive applications, the ambiguities in social media data have made it difficult to parse and interpret meaning, even with state-of-the-art techniques. Hence, researchers have posed conservative attitudes regarding the application of assessment of social media text, such as sensory research, are partially due to this restriction (Piqueras-Fiszman, 2015). Furthermore, due to missing, mislabeled, inaccurate, or potentially spurious information representation, the nature of indigenous unstructured information is often in dirty format (Wang & Jones, 2017). One of the biggest issues affecting data credibility is the resulting low data quality. The lack of large volumes of labeled data also restricts the application of advanced ML and deep learning methods. Data representativity is another issue associated with web mining. For instance, researchers have recently questioned the usage of social media data for inferring health-related outcomes due to the issues of sampling bias (Cesare, Grant, & Nsoesie, 2019; Mooney & Garber, 2019), and effects on validating complex models compared with expert reports (Sandhu, Giabbanelli, & Mago, 2019). Likewise, the integrity of database elements is their accuracy. Data integrity is also a critical issue for database data. For example, authorized users might make mistakes collecting data, computing results, and entering values. Hence, database management systems sometimes have to take action

to catch and correct errors after they are inserted. On the other hand, maintenance of a large number of transactions in a large database is a rather expensive task (Doorn, 2001). As in the example of the blockchain environment, it is necessary to make sure that after data get recorded, they will not be altered (Yli-Huumo, Ko, Choi, Park, & Smolander, 2019).

The privacy issue remains to be a significant concern in cultures when it comes to digitalized information, particularly social media data (Sapienza & Palmirani, 2018). Limited data access and anonymization of sensitive information are two primary methods of protecting privacy (Wu, Zhu, Wu, & Ding, 2014). In addition, only population statistics are revealed without accessing individual identification information. There are often restrictions on accessible information that could be extracted due to API constraints. For instance, less than 1% of Twitter's data, the most common research-based social media platform, could be gathered using its API streaming technique (Vidal et al., 2015). One issue raised by this is that the result may be biased if the sample size is not sufficient. Although paid services such as the Twitter Enterprise API can provide scientists access to all historical data, it is costly to use. Besides social media, the use of other information sources for public health surveillance such as digital records (e.g., EMRs and Fitbit) is also restricted due to privacy issues. As Paul and Dredze (2017) pointed out, consumers may not want their information to be used without asking. When user information is collected, for example, on social media, while the user is unaware, privacy issue can occur. In addition to consumer data, there can be privacy issue on industry sources of data. Kamilaris et al. (2019) discussed the problem in blockchain environment, pointing out that the technology may increase transparency on one hand, while creating privacy issues on the other hand. In food supply systems, privacy is particularly important for keeping one's competitive advantages in the market. Privacy issues would be an ongoing subject of discussion; however, the advantages of using these information sources with the objective of information sharing should not be abandoned entirely. More efforts should be made on improving user privacy while allowing data sharing and utilization.

Furthermore, the lack of computational skills for big data analysis is one of the biggest barriers to the full realization of its potential in other fields (Chiang et al., 2017). Piqueras-Fiszman (2015) presented a similar concern when considering its applications in rapid sensory evaluations. Today, a lot of user-friendly software are being created to assist individuals to collect and analyze information more readily without a computer science background. It is expected that the fast advances in the area of data science would reduce the cost of skill learning and implementation. However, for anyone interested in the technology, fundamental understanding of digital data analysis along with popular tools is crucial.

7 | CONCLUSIONS

This paper summarizes the use of text data in seven areas of research linked to food, nutrition, and health. It is evident that social media has received considerable attention from academia and industry and has been applied widely to almost every aspect of the research mentioned above. The use of social media information has allowed us to "see" hidden patterns via a "big data scope." Nevertheless, broad application usage of social media is restricted by inferior data quality and privacy issues. Other sources of text data, though small, have been used to create useful information systems for helping consumers making decisions on purchasing, cooking, and eating. Text has been the vital media in human communication in any natural language. Our communication is made more efficient either by discovering hidden knowledge from text or by developing text-based information systems. Although several pitfalls remain to be the area for further investigation, we have seen that advanced text mining techniques could help address critical issues in food and nutrition sciences.

ACKNOWLEDGMENT

This study was partially supported by a USDA Specialty Crop Block Grant award through Illinois Department of Agriculture (698 IDOA SC-19-06) and the Illinois Agricultural Experiment Station.

AUTHOR CONTRIBUTIONS

Dandan Tao searched the literature and drafted the manuscript. Dr. Pengkun Yang edited Section 3 and provided suggestions on discussion of technical parts. Dr. Hao Feng reviewed and edited the manuscript.

ORCID

Hao Feng https://orcid.org/0000-0002-1703-2194

REFERENCES

Abbar, S., Mejova, Y., & Weber, I. (2015). You tweet what you eat: Studying food consumption through twitter. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3197–3206), Seoul, Korea.

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Berlin: Springer Science & Business Media.

Ahn, Y. Y., Ahnert, S. E., Bagrow, J. P., & Barabási, A. L. (2011). Flavor network and the principles of food pairing. *Scientific Reports*, *1*, 196. https://doi.org/10.1038/srep00196.

Ahnert, S. E. (2013). Network analysis and data mining in food science: The emergence of computational gastronomy. *Flavour*, *2*(1), 4. https://doi.org/10.1186/2044-7248-2-4.

Aiello, L. M., Schifanella, R., Quercia, D., & Del Prete, L. (2019). Large-scale and high-resolution analysis of food purchases and

health outcomes. *EPJ Data Science*, *8*(1), 14. https://doi.org/10.1140/epjds/s13688-019-0191-y.

Al-Gaadi, K. A., Hassaballa, A. A., Tola, E., Kayad, A. G., Madugundu, R., Alblewi, B., & Assiri, F. (2016). Prediction of potato crop yield using precision agriculture techniques. *PloS One*, *11*(9), e0162219.

Alamsyah, A., & Peranginangin, Y. (2015). Network market analysis using large scale social network conversation of indonesia's fast food industry. *Proceedings of the 2015 3rd International Conference on Information and Communication Technology (ICoICT)* (pp. 327–331). Piscataway, NJ: IEEE.

Anbarkhan, S., Stanier, C., & Sharp, B. (2018). Text mining approach to extract associations between obesity and arabic herbal plants. *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications* (pp. 211–220). Cham: Springer.

Ariyasriwatana, W., & Quiroga, L. M. (2016). A thousand ways to say 'delicious!' categorizing expressions of deliciousness from restaurant reviews on the social network site yelp. *Appetite*, *104*, 18–32.

Asano, Y. M., & Biermann, G. (2019). Rising adoption and retention of meat-free diets in online recipe data. *Nature Sustainability*, *2*(7), 621–627.

Astill, J., Dara, R. A., Campbell, M., Farber, J. M., Fraser, E. D., Sharif, S., & Yada, R. Y. (2019). Transparency in food supply chains: A review of enabling technology solutions. *Trends in Food Science & Technology*, *91*, 240–247.

Bashir, N., Papamichail, K. N., & Malik, K. (2017). Use of social media applications for supporting new product development processes in multinational corporations. *Technological Forecasting and Social Change*, *120*, 176–183.

Bhimani, H., Mention, A. L., & Barlatier, P. J. (2019). Social media and innovation: A systematic literature review and future research directions. *Technological Forecasting and Social Change*, *144*, 251–269.

Blackburn, K. G., Yilmaz, G., & Boyd, R. L. (2018). Food for thought: Exploring how people think and talk about food online. *Appetite*, *123*, 390–401.

Boulos, M. N. K., Yassine, A., Shirmohammadi, S., Namahoot, C. S., & Brückner, M. (2015). Towards an "Internet of Food": Food ontologies for the Internet of Things. *Future Internet*, *7*(4), 372–392.

Bouzembrak, Y., & Marvin, H. J. (2016). Prediction of food fraud type using data from Rapid Alert System for Food and Feed (RASFF) and Bayesian network modelling. *Food Control*, *61*, 180–187.

Bouzembrak, Y., & Marvin, H. J. (2019). Impact of drivers of change, including climatic factors, on the occurrence of chemical food safety hazards in fruits and vegetables: A Bayesian Network approach. *Food Control*, *97*, 67–76.

Bouzembrak, Y., Camenzuli, L., Janssen, E., & Van der Fels-Klerx, H. J. (2018). Application of Bayesian Networks in the development of herbs and spices sampling monitoring system. *Food Control*, *83*, 38–44.

Carr, J., Decreton, L., Qin, W., Rojas, B., Rossochacki, T., & wen Yang, Y. (2015). Social media in product development. *Food Quality and Preference*, *40*, 354–364.

Centers for Disease Control and Prevention (CDC). (2012). *Overweight & obesity*. Retrieved from http://www.cdc.gov/obesity

Centers for Disease Control and Prevention (CDC). (2018). *Annual summaries of foodborne outbreaks*. Atlanta, GA: US Department of Health and Human Services, CDC. Retrieved from https://www.cdc.gov/fdoss/annual-reports/index.html

Çelik, D. (2015). FoodWiki: Ontology-driven mobile safe food consumption system. *Scientific World Journal, 2015*. https://doi.org/10.1155/2015/475410.

Cesare, N., Grant, C., & Nsoesie, E. O. (2019). Understanding demographic bias and representation in social media health data. *Proceedings of the Companion Publication of the 10th ACM Conference on Web Science* (pp. 7–9). New York, NY: ACM.

Chaix, E., Deléger, L., Bossy, R., & Nédellec, C. (2019). Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology*, *81*, 63–75.

Chen, M., Jia, X., Gorbonos, E., Hong, C. T., Yu, X., & Liu, Y. (2019). Eating healthier recipe recommendation. *Information Processing & Management*, 10251.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8*(6), 866–883.

Chen, S., Huang, D., Nong, W., & Kwan, H. S. (2016). Development of a food safety information database for Greater China. *Food Control*, *65*, 54–62.

Chiang, L., Lu, B., & Castillo, I. (2017). Big data analytics in chemical engineering. *Annual Review of Chemical and Biomolecular Engineering*, *8*, 63–85.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creativity and Innovation Management*, *26*(1), 17–30.

Correa, J. C., Garzón, W., Brooker, P., Sakarkar, G., Carranza, S. A., Yunado, L., & Rincón, A. (2019). Evaluation of collaborative consumption of food delivery services through web mining techniques. *Journal of Retailing and Consumer Services*, *46*, 45–50.

De Choudhury, M., Sharma, S., & Kiciman, E. (2016). Characterizing dietary choices, nutrition, and language in food deserts via social media. *Proceedings of the 19th ACM Conference on Computer-supported Cooperative Work & Social Computing* (pp. 1157–1170). New York, NY: ACM.

De Clercq, M., Stock, M., De Baets, B., & Waegeman, W. (2016). Data-driven recipe completion using machine learning methods. *Trends in Food Science & Technology*, *49*, 1–13.

do Nascimento, A. B., Fiates, G. M. R., dos Anjos, A., & Teixeira, E. (2013). Analysis of ingredient lists of commercially available gluten-free and gluten-containing food products using the text mining technique. *International Journal of Food Sciences and Nutrition*, *64*(2), 217–222.

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., … Hsiao, W. W. (2018). FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Science of Food*, *2*(1), 23.

Doorn, J. H. (Ed.). (2001). *Database integrity: Challenges and solutions*. Hershey, PA: IGI Global.

Drury, B., & Roche, M. (2019). A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture*, *163*, 104864.

Du, C. J., & Sun, D. W. (2006). Learning techniques used in computer vision for food quality evaluation: A review. *Journal of Food Engineering*, *72*(1), 39–55.

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

Effland, T., Lawson, A., Balter, S., Devinney, K., Reddy, V., Waechter, H., … Hsu, D. (2018). Discovering foodborne illness in online restaurant reviews. *Journal of the American Medical Informatics Association*, *25*(12), 1586–1592.

Eftimov, T., Ispirova, G., Potočnik, D., Ogrinc, N., & Seljak, B. K. (2019). ISO-FOOD ontology: A formal representation of the knowledge within the domain of isotopes for food science. *Food Chemistry*, *277*, 382–390.

El Bilali, H., & Allahyari, M. S. (2018). Transition toward sustainability in agriculture and food systems: Role of information and communication technologies. *Information Processing in Agriculture*, *5*(4), 456–464.

Ertuğrul, D. Ç. (2016). Foodwiki: A mobile app examines side effects of food additives via semantic web. *Journal of Medical Systems*, *40*(2), 41.

Esteki, M., Regueiro, J., & Simal-Gándara, J. (2019). Tackling fraudsters with global strategies to expose fraud in the food chain. *Comprehensive Reviews in Food Science and Food Safety*, *18*(2), 425–440.

Fardet, A., Lakhssassi, S., & Briffaz, A. (2018). Beyond nutrient-based food indices: A data mining approach to search for a quantitative holistic index reflecting the degree of food processing and including physicochemical properties. *Food & Function*, *9*(1), 561–572.

Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., & Bell, D. (2014). Analyzing the language of food on social media. *Proceedings of the 2014 IEEE International Conference on Big Data (Big Data)* (pp. 778–783), Washington DC.

Fritsche, J. (2018). Recent developments and digital perspectives in food safety and authenticity. *Journal of Agricultural and Food Chemistry*, *66*(29), 7562–7567.

Galletti, A., & Papadimitriou, D. C. (2013). *How big data analytics are perceived as a driver for competitive advantage: A qualitative study on food retailers*, pp. 1–59 (Master's thesis, Uppsala University, Uppsala, Sweden).

Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2017). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality & Tourism*, *18*(4), 465–492

García-León, R. A. (2019). Twitter and Food Well-being: Analysis of #Slowfood Postings Reflecting the Food Well-being of Consumers. *Global Media Journal México*, *16*(30), 91–112.

Ge, M., Elahi, M., Fernaández-Tobías, I., Ricci, F., & Massimo, D. (2015). Using tags and latent factors in a food recommender system. *Proceedings of the 5th International Conference on Digital Health* (pp. 105–112). New York, NY: ACM.

Ghosh, D., & Guha, R. (2013). What are we tweeting about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90–102.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., … Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. *Science*, *327*(5967), 812–818.

Greenfield, H., & Southgate, D. A. (2003). *Food composition data: Production, management, and use*. Rome: FAO.

Greis, N. P., & Nogueira, M. L. (2017). A data-driven approach to food safety surveillance and response. In S. Kennedy (Ed.), *Food protection and security* (pp. 75–99). Amsterdam: Elsevier.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Amsterdam: Elsevier.

Harris, J. K., Mansour, R., Choucair, B., Olson, J., Nissen, C., & Bhatt, J. (2014). Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013–2014. *Morbidity and Mortality Weekly Report*, *63*(32), 681–685.

Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., … Balter, S. (2014). Using online reviews by restaurant patrons to identify unreported cases of foodborne illness-new york city, 2012–2013. *Morbidity and Mortality Weekly Report*, *63*(20), 441–445.

Hayashi, Y., Hsieh, M.-H., & Setiono, R. (2009). Predicting consumer preference for fast-food franchises: A data mining approach. *Journal of the Operational Research Society*, *60*(9), 1221–1229.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472.

Hofmann, T. (2017). Probabilistic latent semantic indexing. *ACM SIGIR Forum*, *51*(2), 211–218.

Hornik, K., & Hornik, M. K. (2018). Package 'NLP'.

Huang, Y., Huang, D., & Nguyen, Q. C. (2019). Census tract food tweets and chronic disease outcomes in the US, 2015–2018. *International Journal of Environmental Research and Public Health*, *16*(6), 975.

Ibrahim, N. F., & Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems*, *121*, 37–50.

Ishangulyyev, R., Kim, S., & Lee, S. H. (2019). Understanding food loss and waste—Why are we losing and wasting food? *Foods*, *8*(8), 297.

Jensen, K., Panagiotou, G., & Kouskoumvekaki, I. (2014). Nutrichem: A systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Research*, *43*(D1), D940–D945.

Jia, S. (2018). Behind the ratings: Text mining of restaurant customers' online reviews. *International Journal of Market Research*, *60*(6), 561–572.

Jiménez-Carvelo, A. M., González-Casado, A., Bagur-González, M. G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity—A review. *Food Research International*, *122*, 25–39.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

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

Kapoor, A. P., & Vij, M. (2018). Technology at the dinner table: Ordering food online through mobile apps. *Journal of Retailing and Consumer Services*, *43*, 342–351.

Karaa, W. B. A., Mannai, M., Dey, N., Ashour, A. S., & Olariu, I. (2016). Gene-disease-food relation extraction from biomedical database. *Proceedings of the International Workshop Soft Computing Applications* (pp. 394–407). Berlin: Springer.

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 (pp. 1709–1718). New York, NY: ACM.

Kim, A. Y., Ha, J. G., Choi, H., & Moon, H. (2018). Automated text analysis based on skip-gram model for food evaluation in predicting consumer acceptance. *Computational Intelligence and Neuroscience*, *2018*. https://doi.org/10.1155/2018/9293437.

Kim, S., Sung, J., Foo, M., Jin, Y.-S., & Kim, P.-J. (2015). Uncovering the nutritional landscape of food. *PloS One*, *10*(3), e0118697.

King, T., Cole, M., Farber, J. M., Eisenbrand, G., Zabaras, D., Fox, E. M., & Hill, J. P. (2017). Food safety for food security: Relationship between global megatrends and developments in food safety. *Trends in Food Science & Technology*, *68*, 160–175.

Kodan, R., Parmar, P., & Pathania, S. (2019). Internet of things for food sector: Status quo and projected potential. *Food Reviews International*, 1–17. https://doi.org/10.1080/87559129.2019.1657442.

Kosior, K. (2019). Social media analytics in food innovation and production: A review. *Proceedings in Food System Dynamics*, 205–219. https://doi.org/https://doi.org/10.18461/pfsd.2019.1921.

Kuttschreuter, M., Rutsaert, P., Hilverda, F., Regan, Á., Barnett, J., & Verbeke, W. (2014). Seeking information about food-related risks: The contribution of social media. *Food Quality and Preference*, *37*, 10–18.

Lee, J., Ghimire, D., & Rho, J. O. (2013). Rough clustering of Korean foods based on adjectives for taste evaluation. *Proceedings of the 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 472–475). Piscataway, NJ: IEEE.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. arXiv preprint cs/0205028.

Lucas Luijckx, N. B., van de Brug, F. J., Leeman, W. R., van der Vossen, J. M., & Cnossen, H. J. (2016). Testing a text mining tool for emerging risk identification. *EFSA Supporting Publications*, *13*(12), 1154E.

Maeda, Y., Kurita, N., & Ikeda, S. (2005). An early warning support system for food safety risks. *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence* (pp. 446–457). Berlin: Springer.

Maharana, A., Cai, K., Hellerstein, J., Hswen, Y., Munsell, M., Staneva, V., … Nsoesie, E. O. (2019). Detecting reports of unsafe foods in consumer product reviews. *JAMIA Open*, *2*(3), 330–338.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100–103.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60), Baltimore, MD.

Marvin, H. J., Janssen, E. M., Bouzembrak, Y., Hendriksen, P. J., & Staats, M. (2017). Big data in food safety: An overview. *Critical Reviews in Food Science and Nutrition*, *57*(11), 2286–2295.

Masson, E., Bubendorff, S., & Fraïssé, C. (2018). Toward new forms of meal sharing? Collective habits and personal diets. *Appetite*, *123*, 108–113.

Massung, S., Geigle, C., & Zhai, C. (2016). Meta: A unified toolkit for text retrieval and analysis. *Proceedings of ACL-2016 System Demonstrations* (pp. 91–96), Berlin.

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. *Proceedings of the 7th ACM Conference on Recommender Systems* (pp. 165–172). New York, NY: ACM.

Meyer, C. H., Hamer, M., Terlau, W., Raithel, J., & Pongratz, P. (2015). Web data mining and social media analysis for better communication in food safety crises. *International Journal on Food System Dynamics*, *6*(3), 129–138.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Minnens, F., Lucas Luijckx, N., & Verbeke, W. (2019). Food supply chain stakeholders' perspectives on sharing information to detect and prevent food integrity issues. *Foods*, *8*(6), 225. https://doi.org/10.3390/foods8060225.

Montgomery, K., Chester, J., Nixon, L., Levy, L., & Dorfman, L. (2019). Big data and the transformation of food and beverage marketing: Undermining efforts to reduce obesity? *Critical Public Health*, *29*(1), 110–117.

Mooney, S. J., & Garber, M. D. (2019). Sampling and sampling frames in big data epidemiology. *Current Epidemiology Reports*, *6*(1), 14–22.

Moore, J. C., Spink, J., & Lipp, M. (2012). Development and application of a database of food ingredient fraud and economically motivated adulteration from 1980 to 2010. *Journal of Food Science*, *77*(4), R118–R126.

Moskowitz, H. R., & Saguy, I. S. (2013). Reinventing the role of consumer research in today's open innovation ecosystem. *Critical Reviews in Food science and Nutrition*, *53*(7), 682–693.

Mostafa, M. M. (2018). Mining and mapping halal food consumers: A geo-located Twitter opinion polarity analysis. *Journal of Food Products Marketing*, *24*(7), 858–879.

Mouritsen, O. G., Edwards-Stuart, R., Ahn, Y.-Y., & Ahnert, S. E. (2017). Data-driven methods for the study of food perception, preparation, consumption, and culture. *Frontiers in ICT*, *4*, 15. https://doi.org/10.3389/fict.2017.00015.

Muninger, M. I., Hammedi, W., & Mahr, D. (2019). The value of social media for innovation: A capability perspective. *Journal of Business Research*, *95*, 116–127.

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

Nyati, U., Rawat, S., Gupta, D., Aggrawal, N., & Arora, A. (2019). Characterize ingredient network for recipe suggestion. *International Journal of Information Technology*, 1–8. https://doi.org/10.1007/s41870-019-00277-y

Ordun, C., Blake, J. W., Rosidi, N., Grigoryan, V., Reffett, C., Aslam, S., … Klenk, J. (2013). Open source health intelligence (OSHINT) for foodborne illness event characterization. *Online Journal of Public Health Informatics*, *5*(1), e128.

Öztürk, Ö., & Özacar, T. (2019). A case study for block-based linked data generation: Recipes as jigsaw puzzles. *Journal of Information Science*. https://doi.org/10.1177/0165551519849518.

Paul, M. J., & Dredze, M. (2017). Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *9*(5), 1–183.

Pearson, D., & Perera, A. (2018). Reducing food waste: A practitioner guide identifying requirements for an integrated social marketing communication campaign. *Social Marketing Quarterly*, *24*(1), 45–57.

Phan, T. T., Muralidhar, S., & Gatica-Perez, D. (2019). Drinks & crowds: Characterizing alcohol consumption through crowdsensing and social media. *Proceedings of the ACM on Interactive,*

Mobile, Wearable and Ubiquitous Technologies, *3*(2), 59. https://doi.org/10.1145/3328930.

Pinel, F., Varshney, L. R., & Bhattacharjya, D. (2015). A culinary computational creativity system. In T. Besold, M. Schorlemmer, & A. Smaill (Eds.), *Computational creativity research: Towards creative machines* (pp. 327–346). Berlin: Springer.

Piqueras-Fiszman, B. (2015). Open-ended questions in sensory testing practice. In J. Delarue, J. B. Lawlor, & M. Rogeaux (Eds.), *Rapid Sensory Profiling Techniques* (pp. 247–267). Amsterdam: Elsevier.

Rakhi, N. K., Tuwani, R., Mukherjee, J., & Bagler, G. (2018). Data-driven analysis of biomedical literature suggests broad-spectrum benefits of culinary herbs and spices. *PloS One*, *13*(5), e0198030.

Rapid Alert System for Food and Feed (RASFF). (2017). *Directorate general for health and consumer protection*. Brussels: European Commission.

Rejeb, A., Keogh, J. G., & Treiblmaier, H. (2019). Leveraging the Internet of things and blockchain technology in supply chain management. *Future Internet*, *11*(7), 161. https://doi.org/10.3390/fi11070161.

Roman, L., Belorio, M., & Gomez, M. (2019). Gluten-free breads: The gap between research and commercial reality. *Comprehensive Reviews in Food Science and Food Safety*, *18*(3), 690–702.

Rong, C., Liu, Z., Huo, N., & Sun, H. (2019). Exploring Chinese dietary habits using recipes extracted from websites. *IEEE Access*, *7*, 24354–24361.

Ropodi, A., Panagou, E., & Nychas, G.-J. (2016). Data mining derived from food analyses using non-invasive/non-destructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines. *Trends in Food Science & Technology*, *50*, 11–25.

Rutsaert, P., Regan, Á., Pieniak, Z., McConnon, Á., Moss, A., Wall, P., & Verbeke, W. (2013). The use of social media in food risk and benefit communication. *Trends in Food Science & Technology*, *30*(1), 84–91.

Sadilek, A., Kautz, H. A., DiPrete, L., Labus, B., Portman, E., Teitel, J., & Silenzio, V. (2016). Deploying nEmesis: Preventing foodborne illness by data mining social media. *Proceedings of the 28th IAAI Conference* (pp. 3982–3990), Phoenix, AZ.

Sandhu, M., Giabbanelli, P. J., & Mago, V. K. (2019). From social media to expert reports: The impact of source selection on automatically validating complex conceptual models of obesity. *Proceedings of the International Conference on Human-Computer Interaction* (pp. 434–452). Cham: Springer.

Sapienza, S., & Palmirani, M. (2018). Emerging data governance issues in big data applications for food safety. *Proceedings of the International Conference on Electronic Government and the Information Systems Perspective* (pp. 221–230). Cham: Springer.

Sharma, S. S., & De Choudhury, M. (2015). Measuring and characterizing nutritional information of food and ingestion content in instagram. *Proceedings of the 24th International Conference on World Wide Web* (pp. 115–116). New York, NY: ACM.

Simas, T., Ficek, M., Diaz-Guilera, A., Obrador, P., & Rodriguez, P. R. (2017). Food-bridging: A new network construction to unveil the principles of cooking. *Frontiers in ICT*, *4*, 14. https://doi.org/10.3389/fict.2017.00014.

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.

Spence, C., Wang, Q. J., & Youssef, J. (2017). Pairing flavours and the temporal order of tasting. *Flavour*, *6*(1), 4.

Spink, J., & Moyer, D. C. (2011). Defining the public health threat of food fraud. *Journal of Food Science*, *76*(9), R157–R163.

Steinberger, R., Pouliquen, B., & Van der Goot, E. (2013). An introduction to the Europe media monitor family of applications. arXiv:1309.5290.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, *48*, 157–169.

Thakur, M., Olafsson, S., Lee, J. S., & Hurburgh, C. R. (2010). Data mining for recognizing patterns in foodborne disease outbreaks. *Journal of Food Engineering*, *97*(2), 213–227.

Tiozzo, B., Pinto, A., Neresini, F., Sbalchiero, S., Parise, N., Ruzza, M., & Ravarotto, L. (2019). Food risk communication: Analysis of the media coverage of food risk on Italian online daily newspapers. *Quality & Quantity*, 53(6), 2843–2866.

Trattner, C., & Elsweiler, D. (2019). What online data say about eating habits. *Nature Sustainability*, *2*(7), 545–546.

Trattner, C., Kusmierczyk, T., & Nørvåg, K. (2019). Investigating and predicting online food recipe upload behavior. *Information Processing & Management*, *56*(3), 654–673.

U.S. Department of Agriculture (USDA). (2019). *Agricultural Research Service. FoodData Central*. Retrieved from https://fdc.nal.usda.gov

Van de Brug, F. J., Luijckx, N. L., Cnossen, H. J., & Houben, G. F. (2014). Early signals for emerging food safety risks: From past cases to future identification. *Food Control*, *39*, 75–86.

Varshney, L. R., Pinel, F., Varshney, K. R., Bhattacharjya, D., Schörgendorfer, A., & Chee, Y. M. (2019). A big data approach to computational creativity: The curious case of Chef Watson. *IBM Journal of Research and Development*, *63*(1), 7–1.

Vidal, L., Ares, G., Machín, L., & Jaeger, S. R. (2015). Using Twitter data for food-related consumer research: A case study on "what people say when tweeting about different eating situations." *Food Quality and Preference*, *45*, 58–69.

Wagner, C., Singer, P., & Strohmaier, M. (2014). The nature and evolution of online food preferences. *EPJ Data Science*, *3*(1), 38.

Waldner, C. (2017). *Big data for infectious diseases surveillance and the potential contribution to the investigation of foodborne disease in Canada*. Winnipeg, Canada: National Collaborating Centre for Infectious Diseases.

Wang, L., & Jones, R. (2017). Big data analytics for disparate data. *American Journal of Intelligent Systems*, *7*(2), 39–46.

West, R., White, R. W., & Horvitz, E. (2013). From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1399–1410), Brazil.

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107.

Xu, X., & Huang, Y. (2019). Restaurant information cues, Diners' expectations, and need for cognition: Experimental studies of online-to-offline mobile food ordering. *Journal of Retailing and Consumer Services*, *51*, 231–241.

Yang, C., Ambayo, H., Baets, B. D., Kolsteren, P., Thanintorn, N., Hawwash, D., … Lachat, C. (2019). An ontology to standardize research output of nutritional epidemiology: From paper-based standards to linked content. *Nutrients*, *11*(6), 1300.

Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., & Jason, D. (2011). Mining biomedical text toward building a quantitative food-disease-gene network. In M. Biba, & F. Xhafa (Eds.), *Learning structure and schemas from documents* (pp. 205–225). Berlin: Springer.

Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology? A systematic review. *PloS One*, *11*(10), e0163477.

Zhai, C., & Massung, S. (2016). *Text data management and analysis: A practical introduction to information retrieval and text mining*. San Rafael, CA: Morgan & Claypool.

Zhao, Z., Li, M., Li, C., Wang, T., Xu, Y., Zhan, Z., … Chen, Y. (2019). Dietary preferences and diabetic risk in China: A large-scale nationwide Internet data based study. *Journal of Diabetes*. https://doi.org/10.1111/1753-0407.12967.

Zhu, X., Huang, I. Y., & Manning, L. (2019). The role of media reporting in food safety governance in China: A dairy case study. *Food Control*, *96*, 165–179.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Tao D, Yang P, Feng H. Utilization of text mining as a big data analysis tool for food science and nutrition. *Compr Rev Food Sci Food Saf*. 2020;19:875–894. https://doi.org/10.1111/1541-4337.12540