



Information Systems Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Event Warnings

Ahmed Abbasi, Jingjing Li, Donald Adjeroh, Marie Abate, Wanhong Zheng

To cite this article:

Ahmed Abbasi, Jingjing Li, Donald Adjeroh, Marie Abate, Wanhong Zheng (2019) Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Event Warnings. Information Systems Research 30(3):1007-1028. <https://doi.org/10.1287/isre.2019.0847>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2019, The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Event Warnings

Ahmed Abbasi,^a Jingjing Li,^a Donald Adjeroh,^b Marie Abate,^c Wanhong Zheng^d

^a Information Technology Area and Center for Business Analytics, McIntire School of Commerce, University of Virginia, Charlottesville, Virginia 22904; ^b Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, West Virginia 26506; ^c Center for Drug & Health Information, Department of Clinical Pharmacy, School of Pharmacy, West Virginia University, Morgantown, West Virginia 26506; ^d School of Medicine, Robert C. Byrd Health Sciences Center, West Virginia University, Morgantown, West Virginia 26505

Contact: abbasi@comm.virginia.edu,  <http://orcid.org/0000-0001-7698-7794> (AA); jl9rf@comm.virginia.edu (JL); don@csee.wvu.edu (DA); mabate@hsc.wvu.edu (MA); wzheng@hsc.wvu.edu (WZ)

Received: January 10, 2016

Revised: June 22, 2017; August 5, 2018

Accepted: November 2, 2018


Published Online in Articles in Advance:
August 29, 2019

<https://doi.org/10.1287/isre.2019.0847>

Copyright: © 2019 The Author(s)

Abstract. With greater impetus on broad postmarket surveillance, the Voice of the Customer (VoC) has emerged as an important source of information for understanding consumer experiences and identifying potential issues. In organizations, risk management groups are increasingly interested in working with their information technology teams to develop robust VoC listening platforms. Two key challenges have impeded success. First, prior work has leveraged diverse sets of channels, adverse event types, and modeling methods, resulting in diverging conclusions regarding the viability and efficacy of various user-generated channels and accompanying modeling methods. Second, many existing detection methods rely on “mention models” that have low detection rates, have high false positives, and lack timeliness. Following the information systems design science approach, in this research note we propose a framework for examining key design elements for VoC listening platforms. As part of our framework, we also develop a novel heuristic-based method for detecting adverse events. We evaluate our framework and method on two large test beds each encompassing millions of tweets, forums postings, and search query logs pertaining to hundreds of adverse events related to the pharmaceutical and automotive industries. The results shed light on the interplay between user-generated channels and event types, as well as the potential for more robust event modeling methods that go beyond basic mention models. Our analysis framework reveals that user-generated content channels can facilitate timelier detection of adverse events: on average, two to three years or earlier than commonly used databases. The inclusion of negative sentiment polarity in the models can further reduce false-positive rates. Additionally, we find social media channels provide higher detection rates but lower precision than do search-based signals. The search and web forum channels are timelier than Twitter. The proposed heuristic-based method attains markedly better results than do existing methods—with earlier detection rates of 50%–80% and far fewer false positives across an array of VoC channels and event types. The heuristic method is also well suited for signal fusion across channels. Our note makes several contributions to research. The results also have important implications for various practitioner groups, including regulatory agencies and risk management teams at product manufacturing firms.

History: Gediminas Adomavicius, Senior Editor; Wolfgang Ketter, Associate Editor.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as “*Information Systems Research*. Copyright © 2019 The Author(s). <https://doi.org/10.1287/isre.2019.0847>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.”

Funding: This work was funded by the Directorate for Computer and Information Science and Engineering and the Division of Information and Intelligent Systems of the National Science Foundation [Grants IIS-1816005, IIS-1816504, IIS-1552860, IIS-1553109, BDS-1636933, IIS-1236970, and IIS-1236983] and by the Division of Computing and Communication Foundations of the National Science Foundation [Grant CCF-1629450].

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/isre.2019.0847>.

Keywords: signal detection • social media • data mining • smart health • predictive analytics • healthcare analytics • healthcare information technologies

1. Introduction

Product-related adverse events can have profound monetary and societal implications in various industry contexts. For instance, adverse pharmaceutical drug reactions are responsible for between 3% and 12% of all hospital admissions (Ritter 2008), resulting in millions of hospitalizations and more than 100,000 deaths annually (Sarker et al. 2015). The pharmaceutical drug Pradaxa alone has caused 9,000 hospitalizations, 1,000 deaths, and \$650 million in lawsuit settlements since 2014 (Thomas 2014, Colella-Walsh 2019). Similarly, in the automotive industry, Toyota recently settled lawsuits totaling \$3.4 billion and \$2.3 billion for inadequate rust protection on their trucks, and the unintended acceleration “sticky pedal” fiasco, respectively (Schweinsberg 2012, Fredericks 2014). Traditional adverse event reporting mechanisms have involved formal reporting systems that feed into online databases. Examples include the Adverse Event Reporting System (FAERS) of the U.S. Food and Drug Administration (FDA) and the National Highway Traffic Safety Administration’s (NHTSA) safety issues database. Such databases constitute an invaluable data source for postmarket surveillance. However, many studies have noted the limitations of overreliance on a single channel—most notably, limited coverage of the broad set of adverse events encountered by a diverse consumer population (Forster et al. 2012).

With the rise of big data analytics (Agarwal and Dhar 2014) and greater impetus on broader post-market surveillance, the Voice of the Customer (VoC) has emerged as an important source of information for understanding consumer experiences and identifying potential issues (Zabin et al. 2011, Boynton 2013). This is partly due to increased quality, volume, and timeliness of available VoC information, which encompasses various user-generated content channels including social media, search queries, consumer reports, etc. One of the biggest VoC use cases remains managing risk (Zabin et al. 2011, Browne 2015). For instance, in 2014, for the first time ever, the FDA received more adverse drug reports from consumers than from healthcare professionals—and the volume of customer search queries was orders of magnitude higher (White et al. 2013). Similarly, social media channels offer great potential for adverse event detection for a myriad of products (Abrahams et al. 2015, Sarker et al. 2015). Relevant stakeholders interested in leveraging VoC include regulatory agencies, product manufacturers, consumer advocacy groups, and financial investment firms. In such organizations, risk management groups are increasingly interested in working with their information technology (IT) teams to develop robust VoC listening platforms (Fenwick et al. 2011) capable of identifying adverse

events faster and more accurately, resulting in favorable economic and humanistic outcomes.

However, two key challenges have impeded the success of VoC listening platforms. First, the existing body of knowledge has leveraged diverse sets of channels, adverse event types, and modeling methods, resulting in varying results and diverging conclusions regarding the viability and efficacy of various online user-generated channels and accompanying modeling methods (Schmidt-Subramanian et al. 2014, Sarker et al. 2015). As Davies (2016, p. 1) noted, “A myriad tools and techniques can be applied to a VoC program. This complicates the tasks of investment prioritization and feedback alignment.” Second, many existing detection methods rely on “mention models” that have low detection rates, have high false positives, and fail to detect adverse events in a timely manner, rendering them less useful in real-world risk management contexts (Adjeroh et al. 2014). Consequently, “a key stumbling block for many VoC initiatives” is the lack of meaningful, actionable insights (Davies 2016, p. 8). Presently, risk management and monitoring groups, and IT teams that support such groups, are lacking guidelines regarding many key questions such as the following (Schmidt-Subramanian et al. 2014, Davies 2016): “Which channels should we be integrating into our listening platform?” “Which types of detection methods are best suited for our event types?” “How can we design listening platforms that are practical and valuable in our monitoring contexts?” There remains a need to examine the efficacy of various VoC channels for IT applications with implications for consumer safety (Agarwal et al. 2010, Abrahams et al. 2015). Furthermore, recent studies have underscored the need for more robust detection methods applied to these channels that can serve as decision aids for monitoring teams. The two main research questions we seek to answer are as follows:

1. How effectively can various VoC channels be used to detect different types of adverse product events using state-of-the-art signal detection methods?
2. What are the relevant interactions between channels, event types, and modeling methods, and what are their implications for the design of VoC listening platforms?

To tackle these questions, following the information systems (IS) design science approach, in this research note we propose a framework for examining key design elements for VoC listening platforms. As part of our framework, we also develop a novel heuristic-based method for detecting adverse events. We evaluate our framework and method on two large test beds, each encompassing millions of tweets, forum postings, and search query logs pertaining to hundreds of adverse events related to the pharmaceutical and automotive

industries. The results shed light on the interplay between user-generated channels and event types, as well as the potential for more robust event modeling methods that go beyond basic mention models. More specifically, the results from our analysis framework reveal that user-generated content channels can facilitate timelier detection of adverse events: on average, two to three years earlier than commonly used regulatory databases. The inclusion of negative sentiment polarity in the models can further reduce false-positive rates across all three channels. Additionally, we find social media channels provide higher detection rates but lower precision than search-based signals. In the context of more explicit/salient events, search and web forum channels are timelier than Twitter. Furthermore, certain event types such as drug-related product recalls are more challenging to detect using user-generated content channels. Whereas most existing mention models detect less than half of all events earlier, with false-positive rates over 75%, the proposed heuristic-based method attains markedly better results—with earlier detection rates of 50%–80% and far fewer false positives across an array of VoC channels and event types. The heuristic method is also well suited for signal fusion across channels.

Our note makes several key contributions to research and practice. We contribute to the emerging IS body of research developing novel analytics capabilities with important business and societal implications (e.g., Shmueli and Koppius 2011, Chen et al. 2012, Bardhan et al. 2015, Brynjolfsson et al. 2016). From a design science perspective, our contributions include a holistic framework for analyzing key design elements pertaining to VoC listening platforms, as well as a novel heuristic-based event modeling method. Our framework unifies and expounds upon insights and key design elements previously examined in a disparate manner, affording opportunities to better understand the interactions between channels, event types, and modeling methods. The proposed event modeling method offers robust detection capabilities that are largely channel and event agnostic, across multiple industry contexts, thereby shifting the detection paradigm away from the status quo, underperforming mention models.

Finally, our research has managerial implications for various practitioner groups. The framework and results can offer guidelines for risk management groups and IT teams tasked with developing and operating VoC listening platforms. By incorporating provisions for key monitoring objectives and constraints such as timeliness, detection, and false-positive rates, the framework is well suited for use by several stakeholders, including regulatory agencies, manufacturing firms, and advocacy groups. Furthermore, the proposed method provides markedly better detection

capabilities, making VoC listening practical and valuable.

2. Proposed Framework

Organizations broadly recognize the importance of listening to VoC, with risk management cited as a primary use case (Zabin et al. 2011, Abrahams et al. 2013). However, the percentage adoption of robust VoC listening platforms and firms' perceived capability maturity of their platforms have both been problematic (Browne et al. 2015). A core issue is that presently, "knowledge of designing, building, integrating, and modifying" effective VoC listening capabilities remains low (Davies 2016, p. 5). There is a need for design frameworks that can bridge the gap between why organizations deploy VoC listening platforms—namely, to integrate appropriate channels and derive valuable insights—and actual outcomes (Zabin et al. 2011).

Design science provides guidelines for the development of IT artifacts, including constructs, models, methods, and instantiations (Hevner et al. 2004). Several prior studies have utilized a design science approach to develop business intelligence and analytics-related IT artifacts, including frameworks, methods, and instantiations (e.g., Lau et al. 2012, Provost et al. 2015). In this note, we employ the design science approach to develop our proposed analysis framework for designing VoC listening platforms.

When creating IT artifacts in the absence of sufficient guidelines, the design science literature suggests that kernel theories may help govern the development process (Gregor and Hevner 2013). VoC listening is about tapping into the "wisdom of the crowds"—the notion that the aggregation of information from external groups can garner better insights (Surowiecki 2005). This idea has been used to fuel "active" participatory approaches such as crowdsourcing, where organizations engage crowds via contests or other incentivized sharing structures. It has also been used as part of "passive" strategies such as opportunistic crowdsensing or big data analytics applied to crowd-generated data for social intelligence (Zeng et al. 2010, Brynjolfsson et al. 2016). Crowds can perform certain types of tasks fairly effectively, including cognitive tasks such as whether a given product will be successful or whether a product has issues (Surowiecki 2005, Sunstein 2006). An important consideration is the amount of task-related information available to the crowd—wise crowds are able to leverage their knowledge, experiences, and intuition (Sunstein 2006). Additionally, crowd wisdom also embodies the following characteristics: diversity of opinion, independence, decentralization, and suitable aggregation mechanisms (Surowiecki 2005). These characteristics of wise crowds highlight

three important implications for VoC listening platforms: (1) impact of tasks and objectives, (2) attributes of VoC channels, and (3) robustness of signal detection methods. Effective VoC listening platforms must carefully consider the design implications of each of these.

1. *Impact of Tasks and Objectives:* The major value proposition of crowd wisdom is that it can facilitate more accurate, timelier insights, leading to better outcomes (Sunstein 2006, White et al. 2013). Utility can be a relative concept, closely related to what the insights are being used for and by whom. For instance, the PredictIt political stock market’s predictions about election outcomes might be used by investors speculating in prediction markets, journalists covering the election, candidate supporters eager to know who might win, and special interest groups looking to get a jump on prospective winners (Surowiecki 2005). VoC listening platform design needs to take into account stakeholder trade-offs based on respective risk tendencies, operational constraints/capabilities, and major objectives.

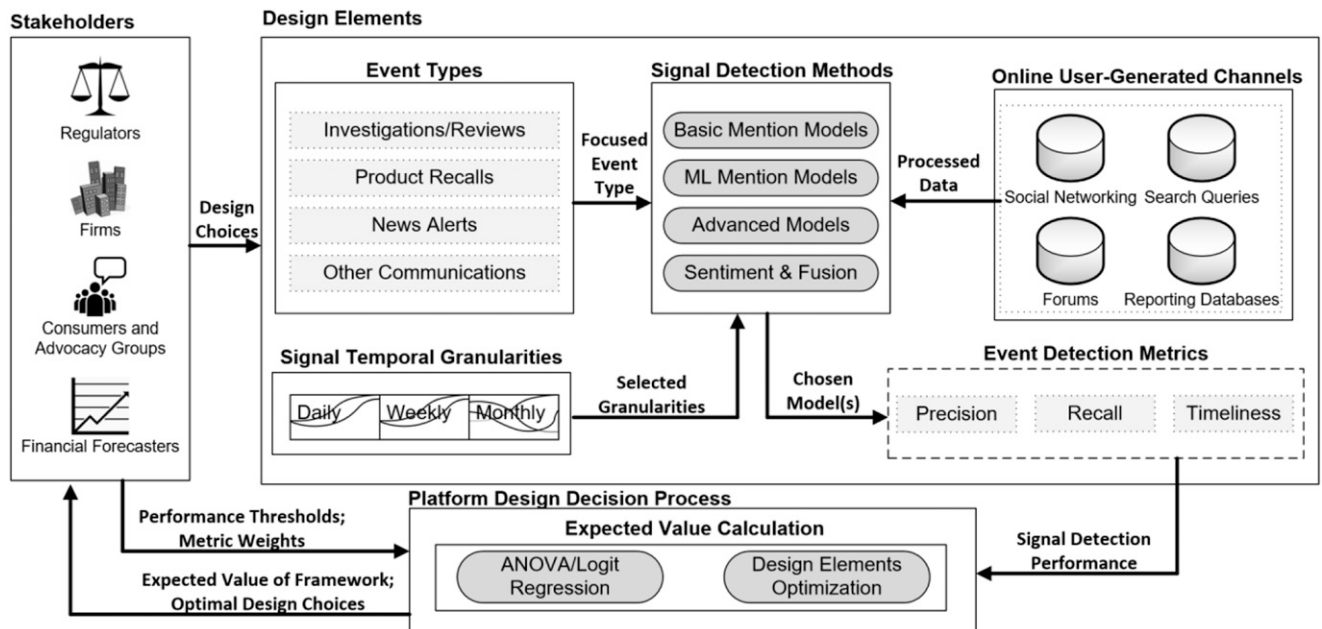
2. *Attributes of VoC Channels:* The diversity, independence, and decentralization of user content generated in VoC channels may have important implications for crowd wisdom-gathering capabilities (Surowiecki 2005). Channels such as search queries and various social media platforms vary in quality, recency, uniqueness, frequency, and salience of content created (Abbasi and Adjeroh 2014). They also encompass varying social network structures that can impact information diffusion patterns (Kwak et al. 2010). Furthermore, these channels may differ in terms of usage

intentions. For instance, search query volume primarily reflects information acquisition patterns (White et al. 2013, Brynjolfsson et al. 2016), whereas forums are used for acquisition, dissemination, and sensemaking via conversations (Abrahams et al. 2015), and Twitter is commonly used for larger-scale broadcasting/dissemination (Kwak et al. 2010). Consequently, VoC listening platform designers must understand cross-channel implications to “prioritize channels based on value” and “justify a more strategic investment” (Davies 2015, p. 6).

3. *Robustness of Signal Detection Methods:* Suitable aggregation mechanisms are essential for effectively leveraging crowd wisdom (Surowiecki 2005). These aggregation mechanisms must perform signal detection, the process of disentangling signal insights from noise (Cassino 2016). As Abrahams et al. (2013, p. 871) note, detecting “whispers of useful information in a howling hurricane of noise” is a huge challenge, and filters are needed to extract meaning from the “blizzard of buzz.” Inadequate signal detection methods can dramatically diminish the utility of VoC listening platforms, and effectively detecting signals from unstructured user-generated channels remains difficult (Browne et al. 2015).

On the basis of these important design implications, we propose a framework for examining VoC listening platforms (depicted in Figure 1). Listening tasks and objectives are represented in the form of stakeholders, event types of interest for monitoring, and the importance of different event detection metrics such as accuracy and timeliness. VoC channels with varying characteristics are represented, including social

Figure 1. Framework for Analyzing VoC Listening Platform Design Elements



media, search queries, and voluntary reporting databases. Key signal detection characteristics incorporated include different types of detection methods and temporal granularities. Collectively, these considerations are incorporated in the platform design decision process, which provides stakeholder-specific insights regarding design choices and their expected values.

Based on Provost and Fawcett (2013) and Blattberg et al. (2008), the design decision process follows three steps: (1) attaining inputs from the stakeholder reflecting their trade-offs associated with model performance outcomes; (2) uncovering the interplay between design elements and model performance; (3) combining stakeholder inputs, model performance, and design elements to prescribe the best design choices that satisfy stakeholder priorities.

Two types of information constitute the necessary stakeholder inputs for the decision process. On the one hand, operations managers may have a different threshold for each metric based on their operational constraints and capabilities—for example, “our team cannot handle more than a certain signal volume, necessitating a higher precision threshold.” We denote these minimum thresholds, which are similar to those in Provost and Fawcett (2013), as mP , mR , and mT . On the other hand, business and risk managers may have varying preferences for these metrics. For example, regulators need to take costly auditing actions for an adverse event, resulting in lower tolerance for false positives. By contrast, firms may be more likely to trade precision for better and timelier recall so that they can proactively cope with adverse events. We denote these preference weights as wP , wR , and wT . These weights are analogous to monetary costs and benefits, reflecting decision maker trade-offs. For example, a higher wP and a lower wR implies that the stakeholder associates a higher cost with false positives than false negatives.

Next, we use analysis of variance (ANOVA) and logit regression to uncover the impact of design elements on signal detection performance metrics

$Y_{ijklm} = \{Precision, Recall, Timeliness\}$. The event type E_k is a between factor (nested under the event type). The online user-generated channel D_i , signal detection method M_j , and temporal granularity T_l are within factors. With S_m standing for individual events, $S/E_{m(k)}$ denoting events within each type, and $\varepsilon_{m(ijkl)}$ as an error term, the structural model describing the sources of variance becomes

$$Y_{ijklm} = Y_t + E_k + S/E_{l(k)} + D_i + M_j + T_l + \varepsilon_{m(ijkl)}. \quad (1)$$

Accordingly, we can predict the possible modeling performance for each of the factorial (design element) combinations by calculating the marginal means, denoted as $\overline{P_{ijklm}}$, $\overline{R_{ijklm}}$, and $\overline{T_{ijklm}}$.

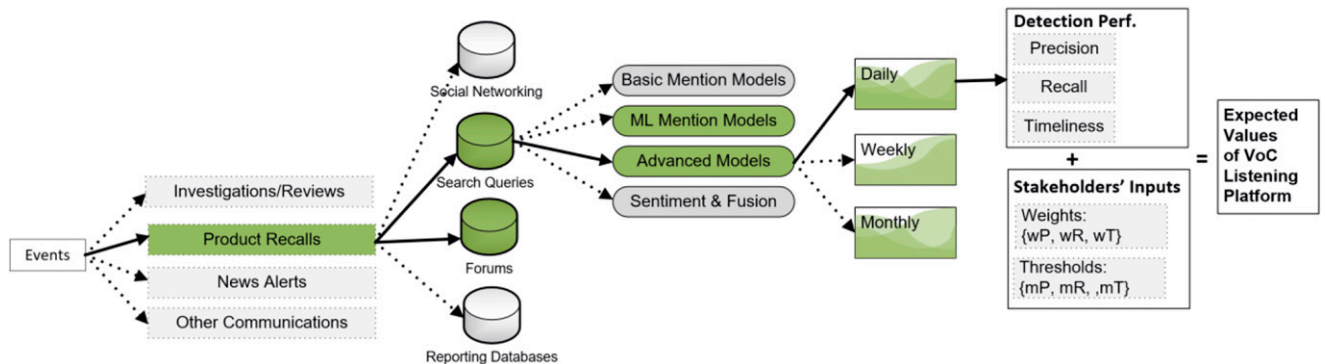
Finally, we generate the expected value of any design combination by considering preference weights and modeling performance (Blattberg et al. 2008, Provost and Fawcett 2013). The design element optimization is formulated as follows:

$$\begin{aligned} \operatorname{argmax}_{\overline{P}, \overline{R}, \overline{T}} (wP \times \overline{P_{ijklm}} + wR \times \overline{R_{ijklm}} + wT \times \overline{T_{ijklm}}) \quad (2) \\ \text{s.t. } \overline{P_{ijklm}} > mP \ \& \ \overline{R_{ijklm}} > mR \ \& \ \overline{T_{ijklm}} > mT, \end{aligned}$$

where $\overline{T_{ijklm}}$ is a linearly transformed timeliness measure ranging between 0 and 1.

Figure 2 illustrates an example of the design decision process (Blattberg et al. 2008). Stakeholders first provide design choices regarding relevant design elements (highlighted in green) specific to their contexts. In this example, the focused event type is *product recalls*; the processed data are from *search queries* and *forums*; the selected temporal granularities are *daily*, *weekly*, and *monthly*; and the chosen models are *machine learning (ML) mention models* and *advanced models*. Within this design choice solution space, signal detection performances for each of the design element combinations can be calculated. Combined with stakeholders’ performance thresholds and metric weights, the platform design decision process can leverage ANOVA/logit regression and design

Figure 2. (Color online) An Example of Platform Design Decision Process



element optimization to identify the optimal design choice leading to the best expected value, which is search query–advanced models–daily (path in solid arrows). At this point, the stakeholder can choose to accept the design, tweak its preference weights, or expand its design choices (e.g., consider alternative VoC channels) to continue to exploring designs with potentially better expected values. We later provide empirical results to demonstrate how the framework can serve as a decision aid for VoC listening platform design. The framework affords opportunities for examining the adverse event detection capabilities of different design configurations, as well as the overall impact of various platform design elements. In the ensuing section, we discuss each component of the framework, including the state of the art, limitations, and key gaps.

3. VoC Listening: Related Work and Research Gaps

3.1. VoC Listening Stakeholders and Types of Adverse Events

VoC listening platforms are relevant to several stakeholders, including regulators, manufacturing firms, consumers and advocacy groups, and investors. In the United States, regulatory agencies such as the FDA, NHTSA, and Consumer Product Safety Commission are actively involved in postmarketing surveillance (Chen et al. 2009, Abrahams et al. 2015, Sarker et al. 2015). Manufacturing firms monitor their products to proactively mitigate risk, including “costs of managing reverse flow of products, disposal costs, restitution costs, and legal and liability costs due to any litigation” as well as indirect costs such as “loss of brand image and erosion of market value” (Hora et al. 2011, p. 766). Similarly, investors/markets are interested in detecting product issues that may impact their stock portfolio (Chen et al. 2009).

The key objectives of VoC listening for risk management are better and faster identification of adverse events (Zabin et al. 2011, Yang et al. 2014). From a big data analytics perspective, these objectives translate into three primary metrics: precision, recall, and timeliness. Precision and recall measure the ability to accurately identify adverse events. *Recall* denotes detection rate, whereas *precision* is a measure of false-positive rate, with implications for “alert fatigue.” *Timeliness* is how much earlier an adverse event can be detected, either in comparison with the point in time when the event transpires or relative to status quo detection methods (Hora et al. 2011). It is a significant VoC listening objective because earlier detection can expedite remedial actions, lessening social and monetary costs. Timely detection allowed Johnson & Johnson to efficiently recall 31 million units of Tylenol

in 1982, and Mattel was able to recall nearly 1 million toys containing lead-based paint in 2007. In both cases, early detection allowed product to be pulled from the supply chain before it adversely impacted consumers (Hora et al. 2011). Conversely, analysis of search query volume data could have allowed one to two years earlier detection of a dangerous adverse drug event causing hyperglycemia—potentially exposing one million fewer consumers to the event (White et al. 2013).

It is important to note that different stakeholders might define “better and faster” differently. For instance, a regulatory agency with a panoramic view of an entire industry encompassing thousands of products might have less bandwidth for false positives than a specific manufacturing firm with a single product line and more available monitoring resources. We underscore this point in our evaluation section by including results from the vantage point of regulators (industry level) and an individual firm. Additionally, we provide a platform design decision process component to consider the trade-offs facing different stakeholders and help them find the best design choices based on their thresholds and preferences, which is described in more detail in Section 3.4.

Past studies have examined adverse events that vary with respect to the product or nature of the event. For instance, some studies have analyzed events pertaining to specific categories of products such as pediatric or cancer medications (Hadzi-Puric and Grmusa 2012). Others have focused on events related to a set of manufacturers, such as Honda, Toyota, and Chevrolet (Abrahams et al. 2015). In the context of postmarketing drug listening, Sarker et al. (2015) observe that most prior studies have examined a maximum of 5–10 products. Other studies have emphasized the importance of examining a wider set of products and event characteristics such as product recalls, safety communications, ongoing reviews, and severe warnings (Hora et al. 2011, Abbasi and Adjeroh 2014). In a broader review of social listening research spanning multiple industries and event types (including manufacturing defects, newspaper complaints, consumer electronic experiences), Abrahams et al. (2013) also noted that most studies had relied on a single event type. *From a VoC listening platform perspective, there remains a need to examine an array of important adverse event types related to multiple stakeholders.*

3.2. VoC Channels

Prior studies have noted the limitations of overreliance on any single data source, including uneven reporting patterns from consumers because of a lack of awareness of that particular channel (Yang et al. 2014, Abrahams et al. 2015). For instance, in the

context of adverse drug events, Xu and Wang (2014) find that FAERS yielded detection precision rates below 2.5%—meaning 39 out of every 40 alerts triggered was a false positive. This is consistent with our own evaluation results presented later in Section 6, in the context of pharmaceutical and automotive events. Inevitably, monitoring teams relying on such data sources must be conservative in their assessments as a result of fewer potential needles in the proverbial haystack.

Relevant alternative user-generated content channels are those encompassing consumer-contributed content (Yang et al. 2014). The most common categories incorporated in past studies are social media such as discussion forums and Twitter (Abrahams et al. 2015, Lardon et al. 2015, Sarker et al. 2015) and search query logs (White et al. 2013). Twitter test bed sizes have ranged from a few thousand tweets containing a specific product name to billions of tweets mentioning an entire category of products (e.g., “cancer drugs”) (Sarker et al. 2015). Discussion forums utilized were primarily consumer or product specific—for instance, the Honda-Tech forum for Honda issues (Abrahams et al. 2013) and health discussion forums such as MedHelp, Drugs.com, and DailyStrength for adverse drug events (Yang et al. 2014, Sarker et al. 2015). Query log data have typically been attained from major search engines such as Google, Bing, or Yahoo!, and they usually include search query frequencies over time (Karimi et al. 2015).

Prior studies have typically focused on a single channel. However, these channels exhibit different characteristics with respect to credibility, frequency, and salience (Agarwal et al. 2010, Abbasi and Adjeroh 2014). For instance, on the one hand, social media channels such as Twitter and certain health forums are prone to spam, resulting in lower credibility (Karimi et al. 2015). On the other hand, forums have lower volume of content than Twitter and search queries but exhibit greater salience—forum postings are capable of incorporating greater background and context than a 140-character tweet and far more relative to a query encompassing a few search terms (Abbasi and Adjeroh 2014). Examining the user journey across multiple channels has become a major area of research with applications in e-commerce and marketing (e.g., customer journey and path-to-purchase) (Song et al. 2014). *Similarly, there is a need to examine the effectiveness of different channels in the context of VoC listening. However, it remains unclear what the trade-offs of the user-generated content channels are with respect to detection rates, false positives, and timeliness of signals* (Sarker et al. 2015). *From a VoC listening platform perspective, the lack of cross-channel studies indicates a paucity of insights that can guide multichannel listening strategies.*

3.3. Signal Detection Modeling Methods

Signal detection methods for identifying adverse events can be broadly grouped into two closely related categories: basic mention models and machine-learning-based mention models (Abrahams et al. 2013, Karimi et al. 2015). Basic mention models consider disproportionality in the occurrence of key product and incident-related tuples relative to overall occurrences of these terms. For instance, in the context of adverse drug event detection, basic mention models typically measure a combination of a drug reference, some reaction-related terms, and, in some cases, anatomy or drug administration-related terms (Adjeroh et al. 2014). An example of this would be, “I have been taking Drug x and began experiencing headaches and pain in my lower back.” Several basic mention models have been proposed that leverage these product–incident co-occurrence values as input. Here, we briefly describe these methods; further details appear in Online Appendix A.

Relative risk (RR) is computed as the relative ratio of observed and baseline counts $p(j|i)/p(j)$, where i and j denote mentions of product and effect, respectively (DuMouchel 1999). One noted limitation of RR is its susceptibility to sampling variability in situations where the observed and baseline counts are both small (Karimi et al. 2015). Proportional reporting ratios (PRR) extends RR by considering the co-occurrence of i and j relative to the occurrence of j in instances without i (Yang et al. 2014). Given that i' represents all product i instances devoid of j , PRR can be interpreted as $p(j|i)/p(j|i')$. Another commonly used mention model is the reporting odds ratio (ROR) (Hadzi-Puric and Grmusa 2012): $p(j|i)/p(j'|i)/p(j|i')/p(j'|i)$. In benchmarking studies on two data sets encompassing adverse drug event reporting system, medication order, and abnormal laboratory result instances, ROR performed comparably in some circumstance, and slightly better in others, relative to comparison methods such as PRR (Liu et al. 2013). Information component (IC) is an information theory-based measure that leverages the mutual information between i and j . IC is used by the World Health Organization, often with better results than other methods (Lindquist 2008). It can be computed as $\log_2(p(j|i)/p(j))$. Basic mention models have also been used in non-health event detection contexts. Abrahams et al. (2012) have developed a “smoke words” method in which terms were weighted based on their occurrence in different types of automotive adverse event mentions.

Other mention models have incorporated supervised or unsupervised machine learning methods that learn patterns involving product and incident terms derived from dictionaries, lexicons, and/or thesauri. For instance, Yang et al. (2013) build linear kernel

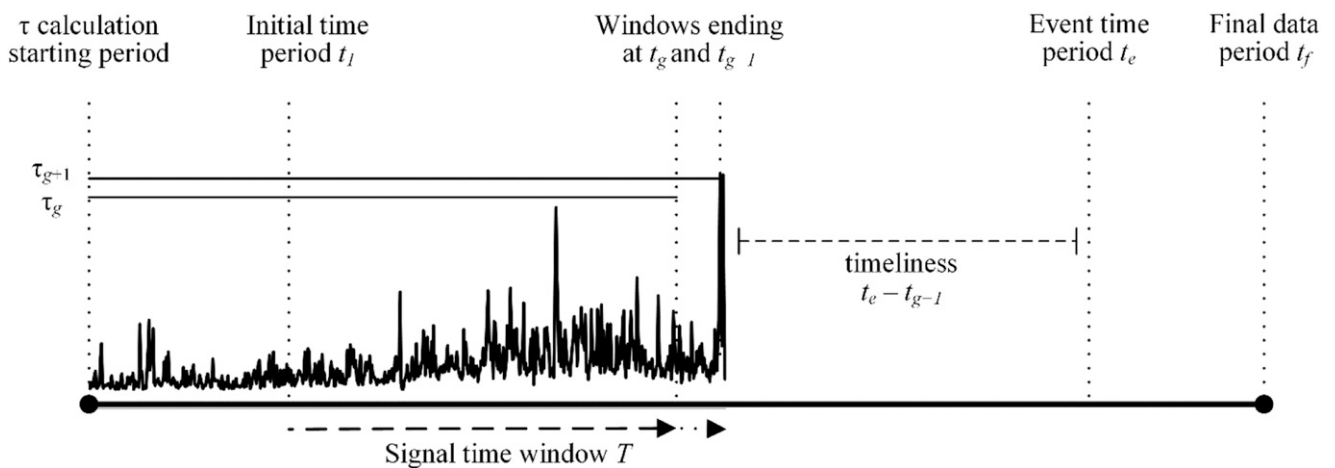
support vector machine (SVM) classifiers that included semantic features such as bag-of-words, product names, and incident term lexicons to determine whether a particular product reference was a valid mention. Abrahams et al. (2013) train linear SVM and naïve Bayes classifiers coupled with a term list feature set to detect vehicle component mentions. Similarly, in the health context, Liu and Chen (2013) develop an SVM classifier that used a custom kernel for relation extraction. For all sentences containing drug and reaction keywords, they derived the shortest dependency path using the StanfordCoreNLP package. Next, they replaced all path tokens with higher-frequency class tokens encompassing part-of-speech tags and entities (e.g., drug, event). The custom kernel function $K(x,y)$ was simply the number of common features between the modified shortest dependency path strings for sentences x and y . SMART combined a logistic regression classifier with a feature set including style, semantic, and product attributes to detect defect mentions in consumer electronics and vehicle discussion forums (Abrahams et al. 2015). Sampathkumar et al. (2014) represented product name, relation keywords, and incidents as hidden states in a hidden Markov model (HMM). Their HMM allowed these three named entities to occur in any order within a message and also included a fourth state for all “other” words within the message. A popular unsupervised method has been association rule mining, where product–incident co-occurrence patterns are derived based on their support and confidence scores (Yang et al. 2014).

Given mention occurrence frequencies over time, time-series analysis can be performed at different temporal granularities (e.g., daily, weekly, monthly, yearly). Similar to prior temporal prediction studies (e.g., Fang et al. 2013), most signal detection methods use windowing to apply temporal association rules or

z-score thresholds to the time series at each time period $t_i \in T = \{t_1, \dots, t_g\}$. This is done by computing these measures over the dynamic time window t_1 to t_g or in some cases beginning with some “training period” to allow suitable thresholds for earlier time periods close to t_1 (Jin et al. 2010, Yang et al. 2014). Figure 3 illustrates how the thresholds τ_g and τ_{g+1} are used for the time-series windows ending at t_g and t_{g+1} , respectively. In this case, the signal at t_{g+1} triggers an alert with timeliness $t_e - t_{g+1}$. To avoid future leaks, for instance, let us assume we are building a monthly model over data beginning in January 2008 (t_1), and we are currently at January 2009 (t_g) in our time-series windowing. The model will only use all data from January 2008 through December 2008 to try to detect a signal. Furthermore, let us assume that this triggers a spike in June 2008. For timeliness purposes, the signal time period will still be considered January 2009 (t_g) because the signal was detected using data up to that point in time. Windowing is performed across all data in the test bed, until t_f .

In summary, existing mention models mostly consider co-occurrence between individual products and incidents. However, many adverse events pertain to product interactions, entire categories of products, and/or incidents encompassing multiple issues. Furthermore, they fail to weight different mention components based on their implications for precision, recall, or timeliness in diverse contexts. Not surprisingly, performance results have varied, with recall rates often below 50% (Adjeroh et al. 2014). Those that examine precision have observed that such methods are prone to high false-positive rates—in many cases, 75% of signals or higher (Adjeroh et al. 2014). *It is unclear how effective existing mention models are when applied to various VoC channels, for a broad array of products and adverse event types. There is a need for more robust signal detection methods beyond basic*

Figure 3. Illustration of Time-Series Setup for Adverse Event Signal Detection



“mention” models, capable of enhanced precision, recall, and timeliness.

3.4. Considering Stakeholder Priorities

Stakeholders have varying priorities based on their risk tendencies, operational constraints/capabilities, and major objectives. These priorities can have a significant impact on the design elements of a VoC listening framework. The expected value framework (Provost and Fawcett 2013) and profitability framework (Blattberg et al. 2008) provide an analytical approach that can guide the design decision process. They begin by decomposing the focal problem into all the possible outcomes of implementing an analytics model, obtain values (e.g., costs and benefits) from stakeholders to factor in trade-offs associated with different outcomes; identify design elements affecting the occurrence probabilities of different outcomes; weight the values by the probabilities of reaching a consolidated expected value; and provide prescriptions about which design element combination and what level of modeling performance (performance thresholds) are needed to accomplish a desirable expected value.

Whereas both frameworks are well suited for guiding evaluations and providing prescriptions for designing VoC platforms, each focuses on binary classification in a business context, necessitating adaptation to our signal detection context for a few reasons. First, for binary classification problems, the confusion matrix determining the possible outcomes is readily attainable. For our signal detection context, however, the possible outcomes worth consideration are beyond the confusion matrix. For instance, for a given event, we only need a single true positive. Additionally, the negatives in signal detection contexts are often harder to understand—an unknown unknown. Second, the temporal aspect of signal detection is critical in practice, which is not well captured in a binary classification context. Finally, in adverse event detection contexts, costs and benefits generally require more time and effort to obtain (e.g., hard to convert the societal benefits and costs into monetary values). *Therefore, it is unclear how effectively existing analytic approaches for guiding design decisions can be adapted to the VoC listening platform context.*

3.5. Inclusion of Sentiment Information and Signal Fusion

Sentiment analysis has seen limited usage in past studies examining adverse events in VoC channels. Sarker and Gonzalez (2015) include sentiment polarity scores derived using the popular SentiWordNet lexicon (Esuli and Sebastiani 2006). They include the overall negative sentiment polarity as an input feature in their SVM classifier and find that the inclusion

of sentiment provided a small lift in mention detection accuracy on their social media data set. Similarly, Yang et al. (2013) include an affect lexicon. Abrahams et al. (2012) use the Harvard General Inquirer dictionary of positive and negative keywords and find that it did not improve vehicle defect identification from online discussion forums. They conclude that general-purpose lexicons might be insufficient to capture nuanced opinion cues appearing in domain-specific online forums. In the context of search, Turney and Littman (2003) propose a simple yet effective method pointwise mutual information method for deriving the sentiment of a search term: by comparing the search query volume for the term plus a set of positively oriented words (e.g., good, positive) and the search query volume for the term and a set of semantically opposed words (e.g., bad, negative).

Similar to sentiment analysis, signal fusion methods have seemingly limited usage for identifying adverse events despite potential for enhancing precision and recall by combining results across channel-specific signals via fusion schemes that are analogous to ensemble voting methods used in meta-learning (Adjeroh et al. 2014). *Given that certain user-generated content channels such as Twitter and search have limited salience (Abbasi and Adjeroh 2014), inclusion of sentiment information could provide an important context refinement regarding user intention in these channels (Sharif et al. 2014). In the same vein, the potential for signal fusion methods to enhance VoC listening capabilities remains underexplored.*

4. Mention Model and Genetic Algorithm-Based Signal Detection

Robust signal detection is essential for effectively tapping into crowd wisdom (Surowiecki 2005). Here, we describe the basic mention model and then discuss our proposed novel heuristic-based method. We use examples related to health adverse drug events, but the methods are generalizable to an array of adverse product event contexts. We later evaluate them on health and automotive test beds.

To identify potential incident references, brand/product, product attribute, and consumer experience lexicons are utilized. An automated tagging tool was developed to assign lexicon tags to references appearing in VoC channel documents. In the health adverse drug event context, these lexicons include drug, anatomy, reaction, and drug administration keywords. For example, the statement “I’ve experienced chest pains ever since I started taking Chantix” would be tagged as “I’ve experienced <ANATOMY> <REACTION> ever since I started taking <DRUG>.” For word-sense disambiguation, we use the CMU part-of-speech tagger designed specifically for short

informal texts to help improve the likelihood that anatomy, side effect, and administration tags were applied appropriately.

4.1. Mention Model for Signal Detection

The basic mention model incorporated in this study can be described as follows. For each product E in our database, we build a fully unsupervised time series. Let $t_k \in T_x = \{t_1, \dots, t_g\}$ signify a given time window, where t_g is the current time period of the analysis and t_g is less than the final time period t_f . Let C_{Dn} represent the number of product names (D) associated with E that appear in a document n . Let $\{d_1, \dots, d_N\}$ signify the set of documents occurring during t_k within a given channel, where each $C_{Dn} \geq 1$. Furthermore, in our health context example, let C_{An} , C_{Rn} , and C_{Mn} represent the accompanying product attribute and customer experience lexicons. These would be number of anatomy (A), reaction (R), and administration (M) terms present in document n , respectively. The aggregated raw score for time t_k is then computed as $s(t_k) = \sum_{n=1}^N C_{Dn} + C_{An} + C_{Rn} + C_{Mn}$. Each $s(t_k)$ is converted to a z-score $z(t_k) = (s(t_k) - \mu_g) / \sigma_g$, where μ_g and σ_g are the mean and standard deviation, respectively, across all t in T_x plus the training period (see Figure 3) where $s(t_k) > 0$. For a given event time series, the basic model considers an alert at time t_k if $z(t_k) > \tau_g$, where τ_g is a threshold for the current window. If t_k is less than the event time period t_e , it is considered a positive signal with timeliness $t_e - t_g$; T_x can vary depending on the resolution of the signals—such as daily or monthly time models, as well as the value of the current window time period t_g .

It is important to note that a single, fixed set of anatomy, administration, and reaction terms are utilized for all drugs. Each product E is represented as a single time series where the y -values are the z-transformed $s(t_k)$. Each event is a spike that exceeds the z-score threshold. Hence, the method is purely unsupervised, without use of any event knowledge a priori. To ensure avoidance of future leaks, neither the drug, reaction, anatomy, and administration terms nor the spikes that are generated use any event information.

4.2. Genetic Algorithm-Based Signal Detection

Effective signal detection entails disentangling signal from noise (Sunstein 2006). One of the biggest limitations of prior mention models has been that they adopt a “one-size-fits-all” approach—applying the same features, weights, and statistical patterns to a diverse set of products, channels, and user experiences. Basic models apply a cookie-cutter disproportionality idea to an array of product–incident mention tuples, resulting in low precision rates. Conversely, supervised

machine learning methods offer better precision but often lack generalizability necessary to garner adequate recall (because of limited diversity of the training data with respect to channels and products). Our proposed genetic algorithm-based signal detection (GASD) method attempts to address these concerns by building signals capable of better accounting for product- and channel-specific characteristics. The two key aspects of the method are its (1) objective function, which rewards the creation of signals that garner fewer, potentially higher-quality, alerts faster; and (2) the weighting method, which allows better contextualization of references to product, attribute, and user experience terms for each individual product. GASD attempts to better harness the diversity of wise crowds for enhanced aggregation, in an unsupervised manner devoid of overfitting. The details are as follows.

GASD learns time-series-specific weights for various product, incident, and experience terms. Extending our drug example, let F_{Dn} represent the occurrence vector of drug terms in document n for product E , and let W_D denote the vector of weights for drug terms where each $W_{Dx} \in \{0, 1/(b-1), 2/(b-1), \dots, 1\}$, and b indicates the number of discrete weight intervals. In GASD, $s(t_k) = \sum_{n=1}^N (\sum W_D F_{Dn} + \sum W_A F_{An} + \sum W_R F_{Rn} + \sum W_M F_{Mn})$, with the objective of finding suitable values for W_D , W_A , W_R , and W_M . Within a population of solutions P , we represent each solution p_q as a binary “bit” string encompassing values for all four sets of terms. Each weight value in p_q is represented using h bits such that there are $2^h = b$ possible weight values for each term. For each p_q , the fitness function $f(p_q)$ is used to evaluate each signal $s(t_k)$ within and across each window T_x . The fitness function considers the timeliness of the signal, the importance of incorporating key reference terms, and provisions to alleviate false positives:

$$f(p_q) = \max_{T_x, k} ((t_f - t_g) D(s(t_k)) (A(s(t_k)) + R(s(t_k)) + M(s(t_k)))) / t_f \ln(l + 1), \quad (3)$$

where t_g denotes the end of a given window T_x , $t_k \in T_x$ indicates one of the time periods that triggers an alert, and $D(s(t_k))$, $R(s(t_k))$, and so forth, indicate the number of drug, reaction, etc., terms appearing in the top r ranked list in period k based on WF values. The variable l denotes the total number of alerts triggered in T_x , used to penalize the fitness value for signals generating excessive alerts. Further details regarding the GASD fitness function and weighting mechanism appear in Online Appendix H.

Figure 4 shows the GASD formulation using the aforementioned fitness function and bit-string encoding. GASD is run for each sliding window instance t_1 to t_g as previously illustrated in Figure 3. For each

Figure 4. GASD Formulation Summary

Given time windows $\{T_1, T_2, \dots\}$, where $T_x = \{t_1, t_2, \dots, t_g\}$

Initialize solution population $P = \{p_1, p_2, \dots, p_y\}$, where $p_q = (p_{q1}, p_{q2}, \dots, p_{qz})$, $p_{qx} \in \{0,1\}$ and the weight of any element in W_D, W_D, W_D , and W_D is represented by $(p_{qx}, p_{qx+1}, \dots, p_{qx+h})$

Repeat until some stopping criterion has been reached

Initialize this generation's solution population $O = \{o_1, o_2, \dots, o_y\}$, where $o_q = ()$

Evaluate each current solution's fitness $f(p_q)$ using formulation in equation (3), across all T_x

Select solutions based on fitness and add to O , where probability of $p_q \in O \propto f(p_q) \left(\sum_{i=1}^y f(p_i) \right)^{-1}$

For each of the $y/2$ solution pairs in O

If random number $u \in [0,1] < c$, crossover o_q and o_{q+1} at point v such that:

$$o_q = (o_q(o_{q1}:o_{qv}), o_{q+1}(o_{q+1v+1}:o_{q+1z})) \text{ and } o_{q+1} = (o_{q+1}(o_{q+11}:o_{q+1v}), o_q(o_{qv+1}:o_{qz}))$$

For each of the y solutions in O

For each o_{qx} in o_q , if random number $u \in [0,1] < m$, mutate o_{qx} such that $o_{qx} = 1 - o_{qx}$

Set O equal to P for the next generation; $P = O$

subsequent generation, the selection probability of a solution is proportional to its $f(p_q)$. Within the new solution set O , crossover is applied on adjacent solutions o_q and o_{q+1} with probability c , and mutation is applied on individual bits within each p_q with probability m . In results reported, $c = 0.7$, $m = 0.001$, and $r = 20$ were used (i.e., no tuning was performed, to avoid overfitting). Stopping criterion for genetic algorithms are an important consideration (Aytug and Koehler 1996). In our analysis, we observed that GASD consistently converged within 200 iterations; however, because run times were not a concern, we used a fixed 500-iteration stopping criterion (i.e., terminate after 500 generations).

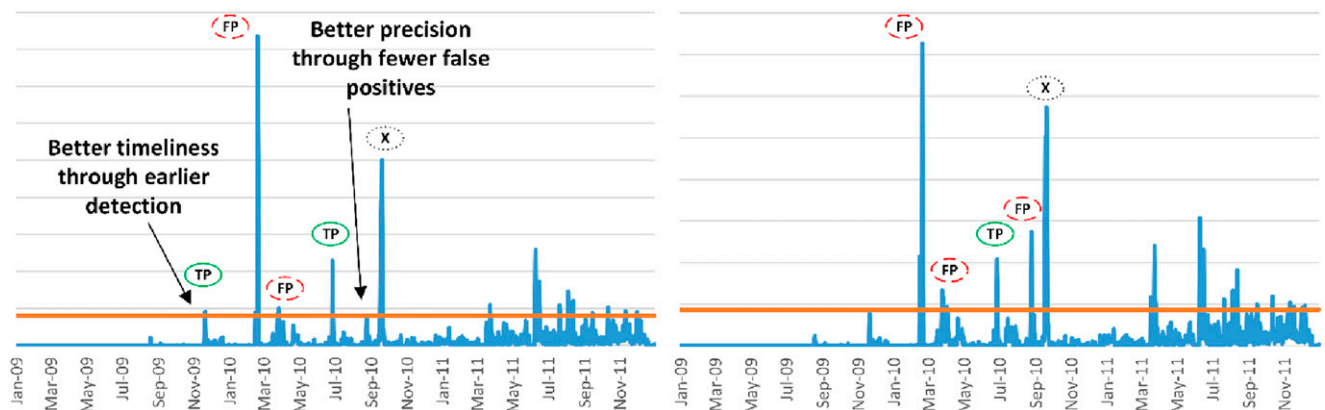
Figure 5 presents a short illustrative example of the effectiveness of GASD. The chart on the left depicts the GASD signal for the drug Actos, relative to the

mention model (right). The FDA announced an investigation on 9/17/2010 for bladder cancer in patients using Actos over an extended period (denoted with an X). The horizontal lines indicate an alert threshold, and TP and FP denote true/false alerts. From the figure it is evident that although both signals appear similar, GASD's term weighting is able to allow earlier detection and fewer false positives by dynamically weighting various drug, reaction, anatomy, and administration keywords, resulting in subtle yet impactful changes in signal strengths.

5. Evaluation Test Bed and Design

Two adverse product event test beds were incorporated. In the main paper, we report results from the health industry, related to adverse drug events. Online Appendix C presents results from the automotive

Figure 5. (Color online) Example Illustrating GASD (Left) vs. Mention Model (Right) Signal for the Drug Actos



industry for adverse automotive events. For our health test bed, we collected data on all drugs that had a *first-time* FDA drug alert for adverse drug events between 2011 and 2013, resulting in 143 events related to 133 unique drugs associated with a myriad of illnesses and ailments, including diabetes, blood pressure, cholesterol, cancer, depression, chronic pain, birth control, insomnia, Parkinson's, arthritis, seizures, etc. The events corresponded to four types: *drug safety communications* are first notifications of a new adverse reaction problem; *ongoing reviews* indicate that the FDA is investigating whether there is a problem; *FDA news* is often forwarded from pharmaceutical company self-reported issues; and *product recalls* are typically due to a manufacturing, packaging, or labeling issue, as opposed to a drug issue. Data from three user-generated content channels were collected: Twitter, forums, and search logs. Table 1 presents an overview. Approximately 12 million tweets containing drug-name keywords spanning 2006 to 2014 were gathered through Topsy's API. Over 5 million postings from 10 popular health forums were attained using web crawlers. The postings spanned the time period 2000 onward. These messages were converted to sentence chunks, resulting in 26 million forum instances in the test bed. This was done because the forum messages were lengthier and often contained discussion of multiple topics. Sentence-level analysis resulted in better performance and information units that were more focused and consistent with the tweets and search queries. Search query frequencies over time were attained from publicly available online sources, as done in prior studies (Brynjolfsson et al. 2016). In particular, we used Google Trends to attain search query volume over time at different temporal granularities for terms in our drug, reaction, anatomy, and administration lexicons, as well as search term co-occurrence volumes. In addition, 6.2 million reports submitted to FAERS were also incorporated in the baseline evaluation to illustrate the value of search and social channels. Consistent with prior studies, for each

report, the set of drugs and reaction terms were used to build the FAERS signals (Xu and Wang 2014).

6. Baseline Evaluation: Comparing Existing Mention Models and Examining Regulatory Databases

Before investigating our core research questions, we conducted two baseline evaluations. The purpose of the first was to illustrate that the baseline mention model utilized in this study was indeed indicative of the types of performance results attained using methods from prior studies. We incorporated several representative comparison methods discussed in the literature review, including basic mention models, machine-learning methods used in prior adverse product event detection studies, and general event detection methods. Performance on the Twitter and forum test beds was examined at three temporal resolutions for the signal time series (day, week, and month). For both channels, we only used data from 2008 onward to allow for better comparison of across-channel performance. For each method, a mention frequency time series was constructed with τ tuned using a grid search. A step interval of 1 was used to compute mean and standard deviation (the basis for the z -score threshold) over a growing window T . For each event, windowing was performed until the FDA first report date.

Consistent with prior work, all methods were evaluated using the standard aforementioned metrics: recall, precision, and timeliness. However, given that our task entails identifying adverse events earlier than the first official mention, "positives" were only those signals that occurred prior to the regulator first-report date for that particular drug event. For each such identified "positive" signal, a determination of true positive (TP) or false positive (FP) was made using a two-stage approach. First, the key drug, reaction, and anatomy keywords appearing in the signal were automatically compared against those appearing in the regulator descriptions. If the similarity was below a certain

Table 1. Overview of Channel and Event Data in Health Test Bed

Channel	Quantity	Time frame	Description
Twitter	12 million tweets	2006–2014	Tweets containing drug keywords
Forums	5 million postings; 26 million sentences	2000–2014	Collected from AskaPatient, Cafepharm, DailyStrength, Drugbuyersguide, Drugs.com, Drugs-Forum, eHealth, MedHelp, MedsChat, PatientsLikeMe
Search	Millions of searches	2004–2014	Query frequency time series aggregated over millions of searches
FAERS	6 million reports	2004–2014	Reports submitted by healthcare professionals and consumers to the FDA Adverse Event Reporting System
Events	Drug safety communications—87; Ongoing reviews—12; Product recalls—35; FDA news alert—9		

threshold, the signal was automatically rejected as a false positive. For those above a threshold, an independent domain expert examined a sample of documents pertaining to the signal (e.g., the underlying tweets, postings, queries) to determine relevance. Precision and recall were computed as $TP / (TP + FP)$ and $(\text{earlier detected events}) / (\text{total events})$, respectively. Timeliness was derived as the earliest average number of days for a detected event, relative to the first FDA event date.

Details regarding the comparison mention models and the evaluation results appear in Online Appendix B. Here, we summarize results for the mention model (labeled “Mention” in Figure 6) and the average results for the three comparison categories of methods evaluated: basic co-occurrence mention models (Basic), machine-learning methods (ML) used in prior adverse product event studies, and general machine-learning methods used in event detection studies (General ML). “Mention” yielded comparable results to those utilized in prior research; it had the highest recall and precision values near the top as well. It also yielded timelier results than other basic co-occurrence models. Overall, the results underscore some of the limitations of existing mention models alluded to in the related work section—they generated low precision (mostly below 20%) and, with the exception of certain daily models, also yielded recall rates below 50%.

As previously alluded to, prior studies have noted the limitations of spontaneous reporting databases such as FAERS (Xu and Wang 2014, Yang et al. 2014). To illustrate the potential of alternative VoC channels,

we ran the four baseline mention models (RR, PRR, ROR, and IC) on the 6 million FAERS reports in our test bed (mentioned in Table 1). Because the FAERS data set did not include text descriptions, only drug and side effect sets, we could not utilize our machine learning methods. We compared the precision, recall, and timeliness of FAERS versus Twitter and forums for the 143 events using these four mention models and found that Twitter and forums yielded significantly better performance across all three performance metrics, for all four mention models (all p -values < 0.001). Figure 7 presents the daily, weekly, and monthly precision, recall, and timeliness results for FAERS, Twitter, and forums, averaged across the four baseline mention models. Consistent with prior studies (e.g., Xu and Wang 2014), FAERS garnered precision rates below 3% and recall rates that were 15–20 points lower than the social media channels. As expected, the timeliness of FAERS true-positive alerts was typically within three to five months of the official first notification date. This is not surprising because FAERS is the primary data source for many of those first notification dates in the first place. Conversely, the social media channels were much timelier.

Our baseline evaluation highlights the potential of alternative VoC channels relative to existing reporting databases, and it shows that the mention model incorporated in this study is representative of prior baseline models. In the following section, we incorporate this mention model as well as the proposed heuristic-based model to examine the effectiveness of user-generated content channels using more robust

Figure 6. (Color online) Performance of Mention Models on Forum (Top Row) and Twitter (Bottom Row) Channels

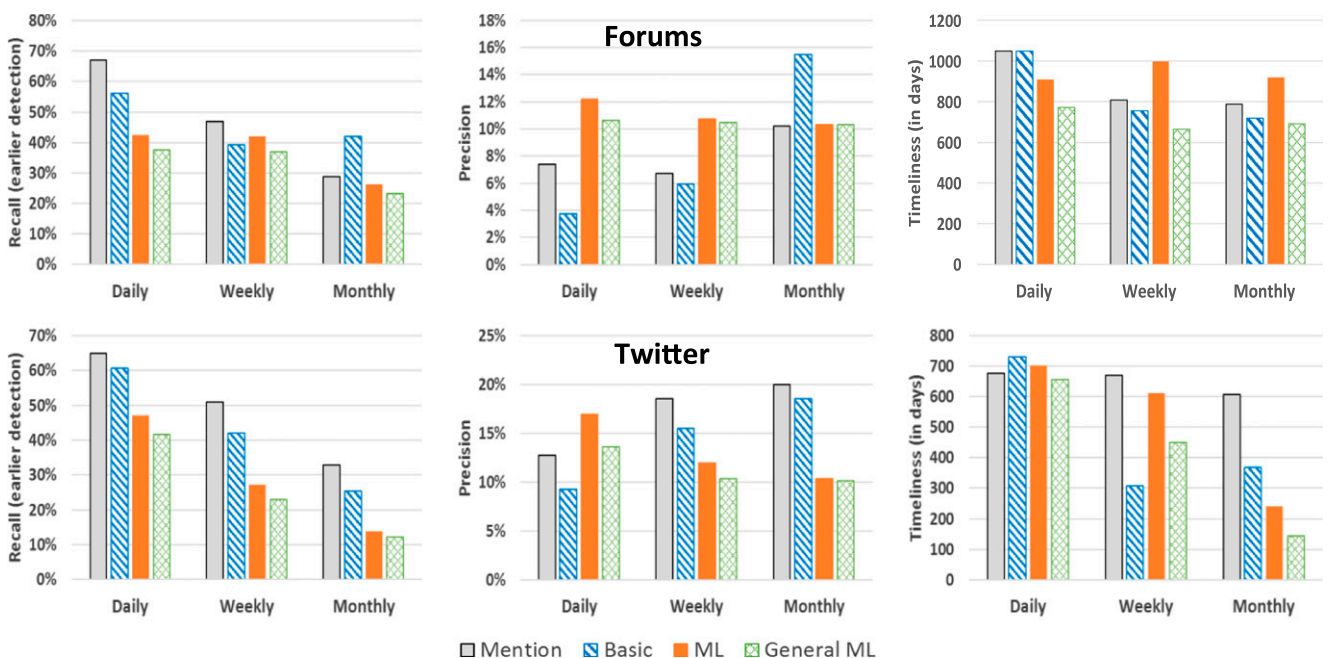
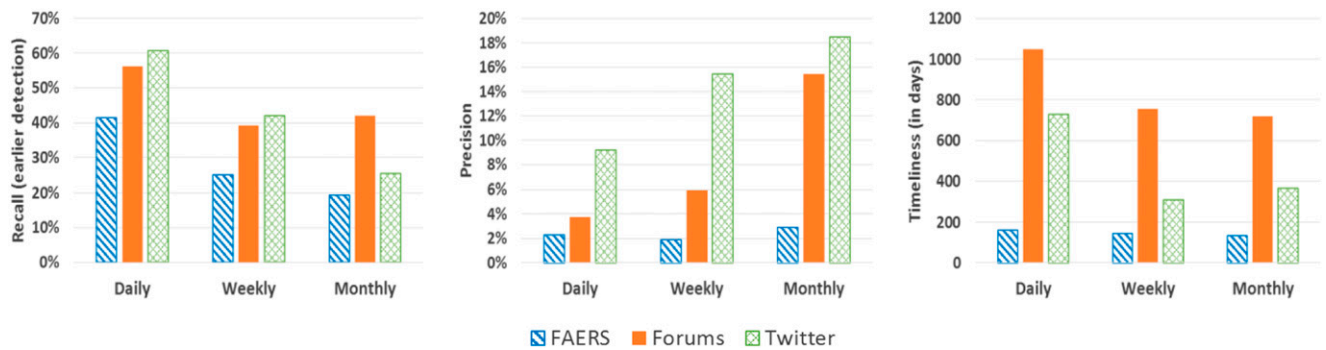


Figure 7. (Color online) Performance of Twitter and Forums Relative to FAERS Using Baseline Mention Models

detection methods, and we assess the interplay between channels, event types, and modeling methods for VoC listening platforms.

7. Evaluation Results

To investigate our two research questions, we used a factorial design encompassing all three channels (Twitter, forums, and search), two types of signal detection models (basic mention model and GASD model), and three temporal granularities for the signal time series (day, week, and month). Initially, sentiment was excluded from the analysis. This resulted in 18 total combinations of signal detection methods. Once again, we only used data from 2008 onward for each channel to allow better comparison of across-channel performance. In all evaluations, $\tau = 1$ was used for GASD as a default (i.e., no tuning), whereas, once again, the best setting for all comparison methods over a grid search was adopted.

7.1. Event Detection Performance

Table 2 presents the results. GASD outperformed the mention model by a wide margin in terms of recall and precision on all settings (typically more than 10 points better). It also yielded timelier signals on most settings, generally detecting events 100–200 days earlier, or more. The recall rates for GASD were in the 55%–80% range, and it attained markedly better precisions rates than prior studies—above 34% for all settings and, in some cases, upwards of 50% or 60%. With respect to channels, forums and Twitter garnered higher recall than search (20 points better on average), but search attained precision rates that were typically at least 10–15 points better. Similarly, the classic trade-off between recall and precision was also observed with respect to temporal resolutions: daily models yielded the best recall but were also prone to the most false positives (likely as a result of greater volatility and noise). Recall rates across event

Table 2. Results Across Channels, Model Types, and Temporal Granularities in Health Test Bed

Channel	Model	Overall metrics			Event type recall			
		Recall	Precision	Time in days mean (SD)	Ongoing review	Safety comm.	Product recall	News alert
Daily models								
Search	Mention	32.2	25.5	1,022 (634)	50.0	31.8	29.4	11.1
	GASD	42.0	48.3	1,132 (489)	50.0	37.6	47.1	0.0
Forums	Mention	67.1	7.4	1,049 (590)	75.0	72.9	52.9	55.6
	GASD	79.7	35.8	1,426 (421)	91.7	80.0	73.5	22.2
Twitter	Mention	65.0	12.7	677 (426)	75.0	75.3	44.1	22.2
	GASD	79.7	35.9	895 (238)	91.7	84.7	73.5	11.1
Weekly models								
Search	Mention	36.4	43.5	1,154 (529)	50.0	43.5	17.6	22.2
	GASD	51.0	56.7	1,183 (421)	58.3	51.8	47.1	11.1
Forums	Mention	46.9	6.7	808 (581)	41.7	54.1	41.2	22.2
	GASD	80.4	33.9	1,338 (435)	91.7	81.2	73.5	33.3
Twitter	Mention	51.0	18.5	669 (408)	75.0	60.0	23.5	22.2
	GASD	74.1	39.1	860 (253)	83.3	80.0	67.6	11.1
Monthly models								
Search	Mention	32.2	42.2	1,075 (538)	33.3	37.6	17.6	11.1
	GASD	51.0	64.4	1,054 (435)	50.0	51.8	47.1	44.4
Forums	Mention	28.7	10.2	787 (570)	25.0	34.1	20.6	11.1
	GASD	77.6	39.4	1,176 (414)	83.3	81.2	67.6	33.3
Twitter	Mention	32.9	20.0	606 (402)	16.7	42.4	17.6	0.0
	GASD	72.7	52.7	777 (267)	83.3	76.5	70.6	22.2

types were generally highest for ongoing reviews and drug safety communications. Not surprisingly, results on product recalls were lower because many of these events are devoid of any explicit reaction or anatomy terms (e.g., “pills chipped or broken in the packaging plant”), making it difficult to detect such signals via VoC channels.

7.2. ANOVA and Logit Regression to Examine Interplay Between Event Types, Channels, and Models

To examine the effect of event types, channels, and model types on our three performance metrics within the 18 model settings, we conducted an ANOVA for each of the 143 events. Specifically, we employed a three-way mixed ANOVA design (i.e., split-plot design). The event type was a between factor (nested under the event type). The channels and models were within factors, indicating that for the same event, signals were repeatedly extracted from different channel and model combinations. We used the weekly granularity models for all analyses to manage the complexity of introducing a fourth variable. Because recall in this setup was binary (i.e., either the event was detected or it was not), we used a logistic mixed model to analyze recall but with the same factorial structure. With S standing for the 143 individual events, A denoting the event type, B signifying channels, and C representing model types, and with S/A denoting events within each type, the structural model describing the sources of variance becomes

$$\begin{aligned}
 Y_{ijkl} = & Y_t + A_k + S/A_{l(k)} + B_i + B \times A_{ik} + B \times S/A_{il(k)} \\
 & + C_j + C \times A_{jk} + C \times S/A_{jl(k)} + B \times C_{ij} \\
 & + A \times B \times C_{ijk} + B \times C \times S/A_{ijl(k)} + \varepsilon_{m(ijkl)}. \quad (4)
 \end{aligned}$$

We do not present the results for precision here because of space constraints; however, the between-factor event type (A) and the within-factors channels (B) and models (C) all had a significant effect on precision (p -values < 0.01). Table 3 depicts the ANOVA results on timeliness. The main effects of event types (A), channels (B), and models (C) were all significant (p -values < 0.01). Interestingly, there was a significant interaction effect between model and channel, as shown in Figure 8(a). Among the three channels, the GASD model obtained the least gain in the search channel and the most gain in the forum channel. Finally, signals detected from the search and forum channels had better timeliness than those detected from Twitter. This latter result has interesting implications that we elaborate on later in Section 8.

As depicted in Table 4, similar to precision and timeliness, event types (A), channels (B), and models (C) all had a significant impact on recall ($p < 0.01$). Additionally, there was a significant interaction effect between event type and channel, as well as channel to model, and a significant three-way interaction among them. Figure 8(b) illustrates the former interaction effect with the channel factor being the comparison basis. As previously alluded to, in general, forums and Twitter were better than search in terms of recall rates. However, the relative strength between forums and Twitter varied depending on event types, with forums performing better on product recalls, whereas the opposite was observed for ongoing reviews.

7.3. Impact of Including Sentiment and Signal Fusion in Mention Model and GASD

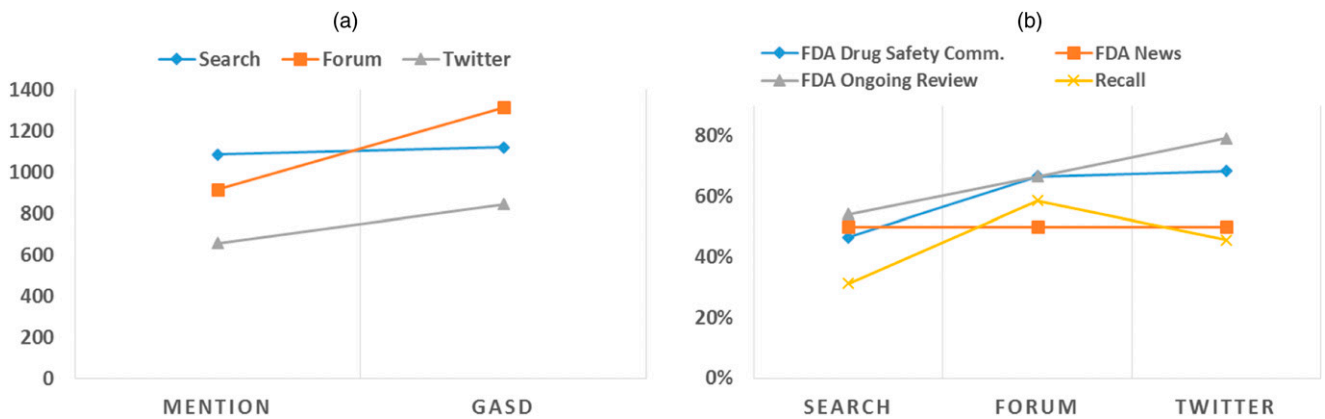
To examine the impact of sentiment, we trained a machine learning sentiment analysis classifier and ran it on the forum and Twitter data sets to derive

Table 3. The ANOVA Results for Timeliness on the “Weekly” Models in Health Test Bed

Source	Partial SS	df	MS	F	Prob $> F$
Model	138,449,409.00	413	335,228.59	5.97	0.000
Between factors					
A	5,211,837.86	3	1,737,279.29	4.33	0.006
S/A	50,906,928.30	127	400,841.96		
Within factors					
B	4,109,855.81	2	2,054,927.90	12.27	0.000
$A \times B$	441,678.48	6	73,613.08	0.44	0.852
$B \times S/A$	26,973,310.20	161	167,536.09		
C	1,542,159.67	1	1,542,159.67	23.24	0.000
$A \times C$	6,685.49	3	2,228.50	0.03	0.992
$C \times S/A$	6,767,840.98	102	66,351.38		
$B \times C$	945,856.44	2	472,928.22	8.42	0.001
$A \times B \times C$	333,606.76	6	55,601.13	0.99	0.438
Residual	4,042,132.29	72	56,140.73		
Total	142,491,541.00	485	293,796.99		

Note. MS, mean square; SS, sum of squares.

Figure 8. (Color online) (a) Interaction Effect Between Channels and Models for Timeliness; (b) Interaction Effect Between Channels and Event Types for Recall



message-level sentiment polarity scores (Hassan et al. 2013, Sharif et al. 2014, Zimbra et al. 2018). For the search channel, the semantic orientation method proposed by Turney and Littman (2003) and previously described in Section 3 was adopted. Details about the sentiment analysis methods utilized appear in Online Appendix D, along with the full evaluation results. Here, we summarize the key takeaways.

Figure 9 shows the precision and recall performance differences for the 18 models with negative sentiment polarity included in the models, relative to the 18 with no sentiment (presented in Table 2). Positive values indicate higher performance with sentiment. Looking at the results, the models with sentiment tended to attain higher precision across the board. However, often the sentiment models also resulted in lower recall. The differences in precision and recall were especially pronounced on the discussion forums, where inclusion of sentiment resulted in marked improvements in precision and decreases in recall. Interestingly, forums are the only channel incorporated that is primarily for the discussion of product issues and experiences. Hence, in this channel, co-occurrence mentions devoid of sentiment information are more likely to result in false positives (Sarker and Gonzalez 2015).

Table 4. The Mixed Logistic Regression Results for Recall on the “Weekly” Models in Health Test Bed

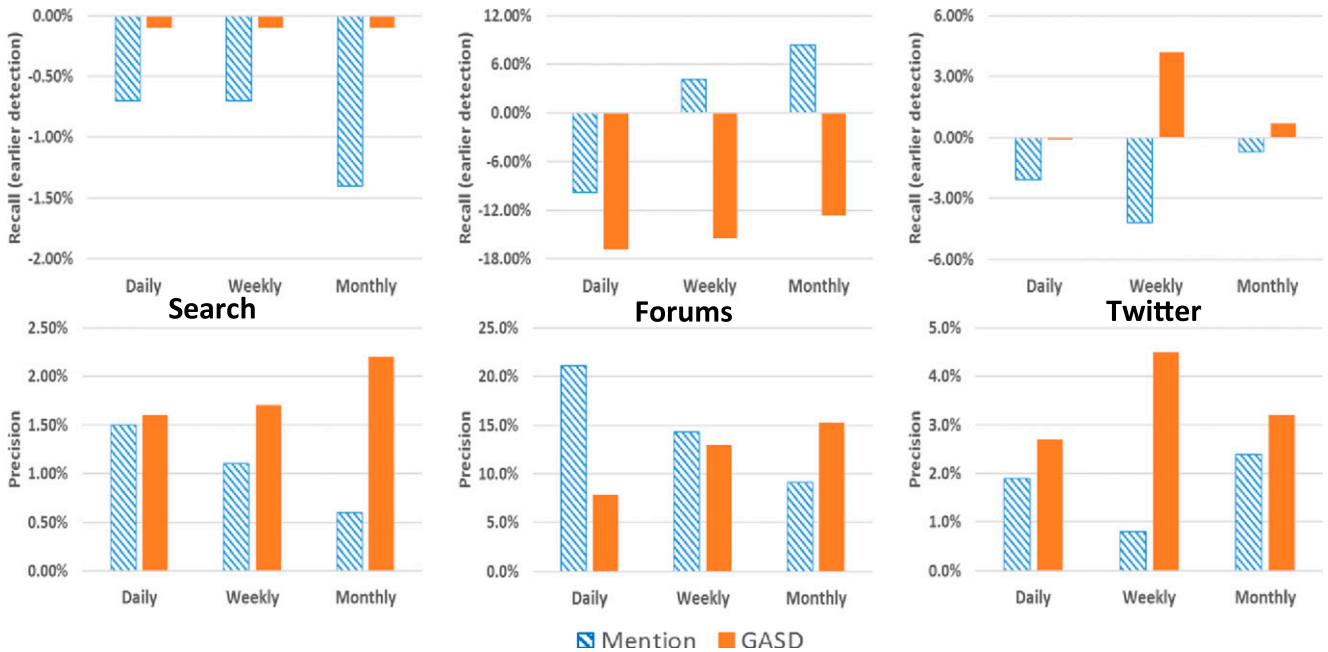
Source	df	χ^2	Prob > χ^2
Between factor			
A	3	22.23	0.000
Within factors			
B	2	110.99	0.000
A × B	6	78.28	0.000
C	1	77.67	0.000
A × C	3	1.86	0.602
B × C	2	11.73	0.003
A × B × C	6	40.82	0.000

Signal fusion was performed by aggregating the individual channels’ signals in a manner analogous to the voting scheme used in ensemble machine learning (Adjeroh et al. 2014). Details about the fusion method utilized appear in Online Appendix E, along with the full evaluation results. Here, we summarize the key takeaways. Figure 10 illustrates the precision, recall, and timeliness performance of the fusion methods on the daily GASD and mention models, relative to the individual channels’ performance. Compared with the three individual channels, fusion improved recall by 3% to 5% versus the best individual channel. Interestingly, precision and timeliness both seemed to move toward the average of the underlying channels incorporated in the fusion. This can be seen in the radar charts, where the fusion line is always “near the middle” on precision and timeliness. As discussed in the appendix, the other interesting observation pertains to recall rates for specific event types. Fusion seemed to garner a greater lift for certain events, such as drug safety communications and FDA news alerts. The results suggest that signal fusion may allow additional degrees of freedom for VoC listening platform stakeholders interested in detecting more events in general, possibly at the cost of less timely and less precise detection or in better detecting specific types of events. Additionally, a subset of channels could also be fused based on detection characteristics that are more conducive to overall monitoring objectives.

7.4. VoC Listening Case Study—Firm Perspective

The results presented in the main paper and Online Appendix C to this point are largely from the perspective of industry regulators such as the FDA and NHTSA tasked with examining adverse events related to a broad array of products spanning multiple firms. However, VoC listening platform stakeholders may include other groups such as individual firms. To examine the efficacy of our framework and proposed

Figure 9. (Color online) Precision and Recall Performance Deltas When Including Sentiment



GASD method from a manufacturer’s vantage point, we present a brief case study from the perspective of the risk management group at Pfizer. We analyzed 20 products from their portfolio, some of which had adverse events that transpired during the time period between 2011 and 2013 (note that some drugs had no events). The events were associated with two types: drug safety communications and product recalls. We ran the mention and GASD models on all Twitter, search, and forum channel data in our test bed and computed precision, recall, and timeliness.

Table 5 shows the evaluation results. GASD was able to detect 76%–84% of the events three to four years earlier. Risk management groups at such firms are often willing to have slightly lower precision (i.e., 33%–60%) for better, timelier recall. Given the size of their monitoring team, and the relatively fewer

products needing monitoring compared with a regulatory agency, such groups are well suited to investigate one or two false alerts for each true positive—a far better ratio than the mention model. Although not depicted, GASD again had lower standard deviations on timeliness.

Figure 11 illustrates the value of the enhanced precision, recall, and timeliness enabled by GASD relative to mention models. Depicted are two Pfizer drugs for which an FDA event transpired. For each event, all GASD and mention model alerts are displayed (true and false positives). For example, an adverse event for the drug Revatio was first detected by GASD 22 months prior to the FDA announcement. In total, GASD had four true positive and two false positive alerts for this product. The figure highlights how signal quality with respect to precision, recall, and

Figure 10. (Color online) Performance Trade-off Implications of Different Channels and Fusion

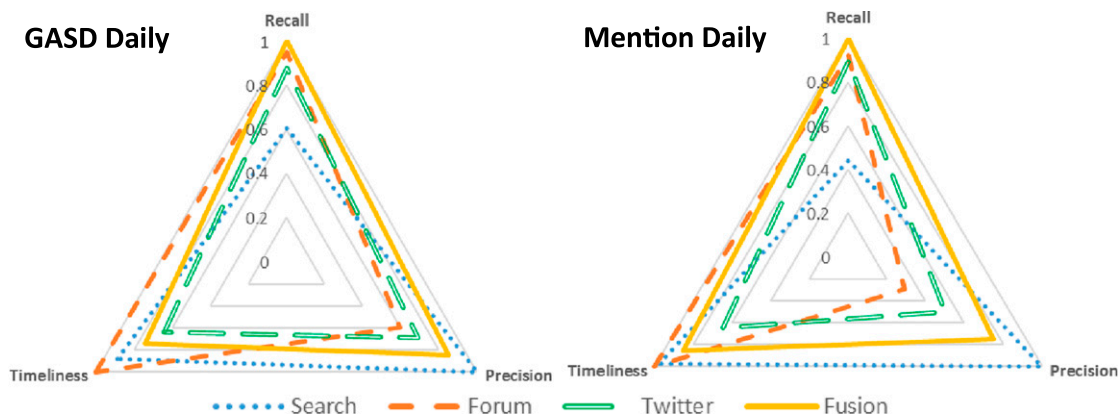


Table 5. Results Across Channels, Model Types, and Temporal Granularities—Pfizer Events

Channel	Model	Daily models			Weekly models			Monthly models		
		Recall (%)	Precision (%)	Time (days)	Recall (%)	Precision (%)	Time (days)	Recall (%)	Precision (%)	Time (days)
Search	Mention	30.8	22.6	1,502	30.8	34.1	1,519	30.8	31.2	1,350
	GASD	30.8	31.0	1,555	30.8	32.7	1,519	30.8	33.5	1,350
Forums	Mention	53.8	4.7	1,182	30.8	2.4	718	15.4	20.0	1,185
	GASD	76.9	33.3	1,454	84.6	27.3	1,407	76.9	32.9	1,080
Twitter	Mention	61.5	6.3	780	46.2	16.7	721	30.8	23.1	192
	GASD	84.6	27.4	914	76.9	35.8	903	61.5	60.3	750

timeliness can impact practical value in real-world settings. Relative to GASD, not only does the mention model fail to detect the Revatio event but also it detects the Zithromax event 18 months after GASD. Furthermore, it generates more false-positive signals, which can cause “alert fatigue” over time, impacting the perceived usefulness of VoC listening capabilities.

7.5. Design Decision Process for Different Stakeholders

As alluded to, the proposed framework can also help different stakeholders to identify the best VoC listening platform design element combination based on their specific thresholds and preferences. Table 6 presents an illustration from the health test bed involving the FDA and Pfizer. In our example, although both stakeholders have similar threshold requirements (i.e., minimum precision, recall, and timeliness), they have differing preference weights for performance metrics. As a regulator, the FDA may have more stringent requirements for precision because of the hefty cost of investigating many signals. Conversely, Pfizer may place greater weight on recall and timeliness to proactively address as many adverse events as possible for monetary and risk mitigation reasons.

We ran ANOVA and logit regression on our two stakeholder’s event data sets, respectively, and we estimated marginal means for all the design element combinations (as described in Section 2). The results are depicted in Table 6. We find that for most (but not all) event types, the FDA should use GASD at weekly or monthly temporal granularities on different channels. By contrast, for the same event types, the best alternative for Pfizer is GASD running on the forum channel at daily intervals. The example indicates how our framework can take stakeholder inputs to determine the best design elements for signal detection. Depending on stakeholder constraints, some event types (e.g., FDA ongoing news in the fourth row) may not be viable for listening under the current thresholds. The example illustrates how our proposed design decision process can help stakeholders identify design elements based on their specific requirements.

8. Results Discussion and Conclusions

Collectively, the evaluation results demonstrate that event types, channels, and models heavily impact event signal detection performance. Table 7 summarizes some of the key results pertaining to our two research questions. In general, GASD provided better

Figure 11. Twitter True/False Positives for Monthly GASD and Mention Models on Two Example Pfizer Drugs (Dark Boxes Denote True Positives; Light Boxes Indicate False Positives)

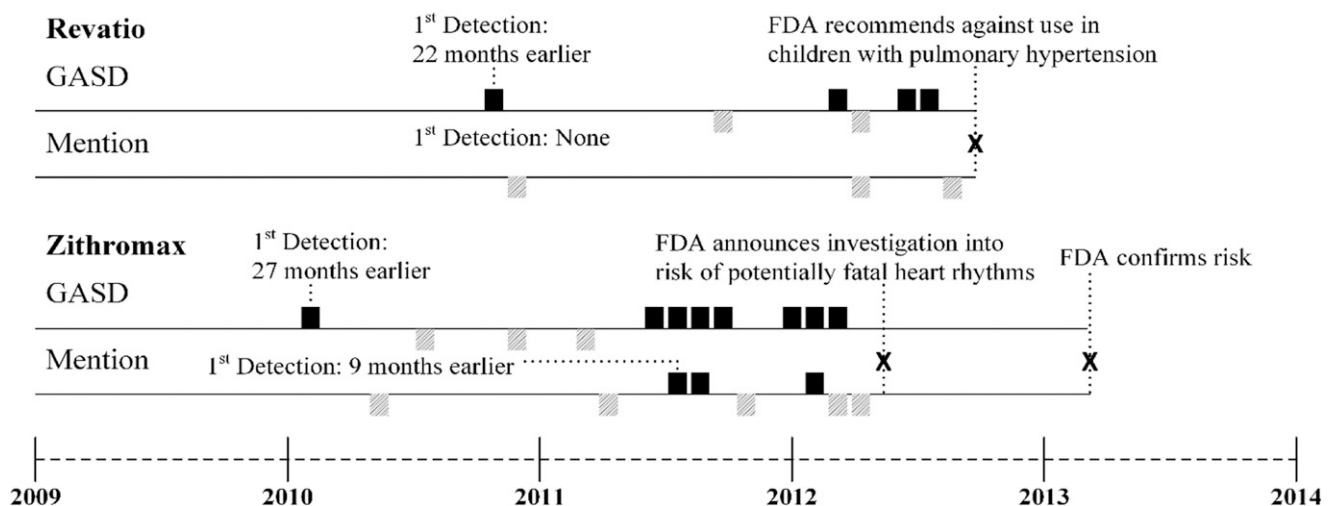


Table 6. Optimal Design Elements for FDA and Pfizer

Stakeholder	<i>mP</i>	<i>mR</i>	<i>mT</i>	<i>wP</i>	<i>wR</i>	<i>wT</i>	Event type	Data/method/time	Prec.	Rec.	Time	Time/	AVG	>T
FDA	0.3	0.2	1,000	0.6	0.3	0.1	Drug safety	Forum/GASD/month	0.43	0.81	1,172	0.68	0.59	Y
	0.3	0.2	1,000	0.6	0.3	0.1	Prod. recall	Forum/GASD/week	0.40	0.74	1,590	0.95	0.58	Y
	0.3	0.2	1,000	0.6	0.3	0.1	FDA news	Search/mention/week	0.50	0.22	1,099	0.64	0.44	Y
Pfizer	0.3	0.2	1,000	0.6	0.3	0.1	Ongoing rev.	Twitter/GASD/month	0.92	0.83	363	0.16	0.84	N
	0.3	0.2	1,000	0.2	0.4	0.4	Drug safety	Forum/GASD/day	0.35	1.00	1,454	0.67	0.73	Y
	0.3	0.2	1,000	0.2	0.4	0.4	Prod. recall	Forum/GASD/week	0.47	0.47	1,774	0.84	0.62	Y

precision, recall, and timeliness than the mention model. GASD’s false-positive rates were two to five times lower than the mention model. The earlier detection rates for many events, across models and user-generated channels (i.e., often one to three years earlier), is also an interesting result that is consistent with some recent studies (White et al. 2013, Adjero et al. 2014). Performance was best on event types with greater salience. In the health test bed, examples included reviews and safety communication events, relative to product recalls. Social media channels provided higher recall but lower precision than did search. Within social media, forums and Twitter

each performed better on certain event types (e.g., forums attained higher recall for product recall events in both the health and automotive test beds). Forums and search also yielded timelier detection than did Twitter.

8.1. Contributions to Research

We contribute to the emerging IS research on application of big data analytics to problems with societal implications (e.g., Chen et al. 2012, Bardhan et al. 2015, Abbasi et al. 2016, Brynjolfsson et al. 2016). From a design science perspective, our contributions include our analysis framework and the GASD

Table 7. Summary of Key Results for Research Questions

Research question	Key results across health and automotive test beds
How effectively can various VoC channels be used to detect different types of adverse product events using state-of-the-art signal detection methods?	<p>User-generated content channels can enable timelier detection of events for 50%–80% of events examined.</p> <p>Using search/social channels, events can be detected one to three years earlier than first reports in official event databases used by regulators and manufacturing firms. Whereas existing mention models suffer from low precision rates (i.e., < 20% on most channels) when applied to search/social, the proposed GASD method garnered false-positive rates that were two to five times lower than those of the mention models, with precision as high as 50%.</p> <p>The inclusion of sentiment in the models enhanced precision, particularly in the online forums. Given that much of the discussion in these forums encompasses incident-related keywords, the extent of negative sentiment may constitute an important filtering mechanism. Similarly, signal fusion can further enhance detection recall and overall <i>f</i>-measures.</p> <p>The aforementioned results with respect to recall (i.e., detection rates), timeliness, and precision (i.e., percentage non-false alarms) were consistent across daily, weekly, and monthly models in test beds spanning two different industry contexts.</p>
What are the relevant interactions between channels, event types, and modeling methods, and what are their implications for the design of VoC listening platforms?	<p>With respect to event types, performance was best on events where users can explicitly mention adverse experiences. In the health test bed, examples include ongoing reviews and drug safety communications. Conversely, drug recalls, which often stem from manufacturing, packaging, or labeling issues, were generally more challenging to detect with search/social channels.</p> <p>Social media channels examined (i.e., forums and Twitter) garnered higher recall but lower precision relative to search. Hence, social media has greater signal and noise, possibly because of the competing effects of greater salience and contextualization on one hand and credibility implications on the other (e.g., spam).</p> <p>Forums and search were timelier than Twitter. This is counterintuitive with findings in other domains such as financial services. The IS literature on crowd-generated data, user motivations for sharing, and online privacy may offer alternative explanations for this effect.</p> <p>Between the two social media channels, forums were better at detecting product recall events in the health and automotive test beds, whereas Twitter provided better detection capabilities for ongoing reviews in the health test bed.</p> <p>The proposed framework can prescribe the best VoC listening platform design choices for a given set of inputs reflecting the stakeholder’s performance trade-offs.</p>

method. The existing literature on adverse event detection has been disparate and fragmented. By using the crowd-generated data literature as a kernel theory, our framework provides a holistic lens for examining key considerations related to VoC listening. An extensive evaluation across two large test beds demonstrated the utility and robustness of the insights generated by our framework and of the GASD method versus existing basic, machine learning, and general-purpose methods. Moreover, the framework can be used to prescribe VoC listening platform design configurations for a given set of stakeholder inputs related to performance trade-offs and constraints.

IS scholars have framed the differences between studies examining prediction versus explanation and the importance of both (Shmueli and Koppius 2011, Agarwal and Dhar 2014, Bardhan et al. 2015). Our analysis framework and the empirical insights produced also provide opportunities for future theory development that offers rich explanations. Using Gregor and Hevner's (2013) classification, our work makes contributions to "nascent theory" that is not yet fully emerged. Three examples are as follows.

1. *Timeliness and Channel Usage Motivations*: Our findings regarding the timeliness of search and forums relative to Twitter are interesting because prior studies on topics such as the relation between social media and stock performance found Twitter to be a stronger lead indicator than forums (Das and Chen 2007, Bollen et al. 2011). As noted in Section 2, the crowd-generated data literature suggest that channel usage intentions (e.g., information acquisition, discussion, and dissemination) might explain the timeliness differences. From a "customer knowledge acquisition journey" perspective, it is possible that internet users sequentially use search to acquire, forums to discuss, and Twitter to disseminate. Alternatively, the earlier use of search and forums, at least in the health test bed, could be attributable to the relatively sensitive nature of the health domain (Anderson and Agarwal 2011). Search constitutes a more private information-gathering option. However, a similar effect was observed in the automotive test bed, a seemingly less sensitive domain. The sharing literature notes that peoples' motives for sharing information and insights may be driven by a desire to help others (Fichman et al. 2011), or for social capital (Wasko and Faraj 2005), and forums provide a more conducive channel for sharing such information compared with Twitter. Our findings suggest an opportunity for studies exploring how certain segments of the population prefer to discuss or disclose their adverse experiences, over time.

2. *Recall and Salience of Different Event Types and Channels*: Our product recall event type had low detection rates in the health test bed. However, in the

automotive context, these types of events had the highest detection rate, likely because of the abundance of sensory and diagnostic cues. For instance, customer mentions included references to sounds, smells, vibrations, and warning lights that are more salient and easily connectable to events. Conversely, dosage errors at a drug bottling plant are far less likely to yield high-quality mentions. The crowd-generated data literature has noted that wise crowds must have sufficient information and knowledge to provide quality insights (Surowiecki 2005, Sunstein 2006). Future research could examine how the amount of knowledge and available information impacts users' quantity and quality of contributions to VoC channels.

3. *Precision of Crowd-Generated Data*: The viability of user-generated content channels comes with caveats. Willingness to disclose information via social media channels, as well as access to and usage of such channels in general, could result in signal sampling biases (Anderson and Agarwal 2011, Abrahams et al. 2015). The social media channels examined in this study garnered higher recall but lower precision relative to search. Hence, they embody greater signal and noise as a result of the competing effects of greater salience on one hand and credibility implications on the other. Furthermore, because certain channels such as search and Twitter may have a degree of separation from event detection tasks in terms of their primary use cases, sole reliance on such channels would not be prudent, as noted by recent issues with the Google Flu monitoring system (Agarwal and Dhar 2014, Lazer et al. 2014). However, the crowd-generated data literature emphasizes the importance of having robust signal aggregation methods (Surowiecki 2005), and other studies have observed that these issues are preventable by including appropriate contextualization mechanisms (Broniatowski et al. 2014, Brynjolfsson et al. 2016). Our study touched upon the potential for signal fusion. An important future direction is to explore more precise signal fusion methods (Adjeroh et al. 2014). Crowdsourcing methods that can enhance signal-to-noise ratios may also offer enhanced detection capabilities (Brynjolfsson et al. 2016). Nevertheless, the results presented in this note constitute an important first step in understanding adverse event detection via crowd-generated data.

8.2. Contributions to Practice

Risk management groups and IT departments can use the analysis framework and empirical insights to develop their VoC listening platforms in a more rigorous, systematic manner (Abbasi et al. 2018, Kitchens et al. 2018). Our framework suggests that practitioners begin by understanding their key monitoring objectives, including which event types they wish to

monitor. The ensuing channel and detection methods investment decisions can be driven by the nuanced precision, recall, and timeliness implications of their environment. For instance, we noted earlier how differences in the quantity of products being monitored and available monitoring resources could result in varying perspectives on precision, recall, timeliness trade-offs for regulators versus individual firms. Furthermore, the results of our GASD method suggest that robust event detection methods applied to appropriate channels have the potential to offer timely detection with manageable false-positive rates, making enterprise VoC listening feasible. Collectively, by addressing many of the key impediments to VoC listening platform adoption and business value (Browne et al. 2015, Davies 2015), this study has the potential to enhance outcomes related to practitioner's VoC listening platform investment decisions.

8.3. Limitations

Our work is not without its limitations. Timeliness is a relative construct: as the status quo changes and the state of the art advances, what is considered timely today may not be in the future. Although omnichannel VoC listening platforms are intended to alleviate some of the availability biases inherent in spontaneous reporting databases, they are not entirely devoid of such biases. Certain products might be better suited to online monitoring. Moreover, disparities such as literacy and socioeconomic factors could moderate the frequency and salience of crowd-generated signals. Consequently, it is conceivable that precision, recall, and timeliness could vary across firms or classes of products, creating potential inequities. From an ethical standpoint, examining adverse event detection biases attributable to, or amplified by, the use of machine learning approaches applied to user-generated content constitutes an important future research direction. Similarly, a deeper exposition into multichannel listening strategies that examine broader stakeholder scenarios and consider additional fusion methods and channels constitutes an important future direction.

References

- Abbasi A, Adjero D (2014) Social media analytics for smart health. *IEEE Intelligence Systems* 29(2):60–64.
- Abbasi A, Sarker S, Chiang RH (2016) Big data research in information systems: Toward an inclusive research agenda. *J. Assoc. Inform. Systems* 17(2):Article 3.
- Abbasi A, Zhou Y, Deng S, Zhang P (2018) Text analytics to support sense-making in social media: A language-action perspective. *Management Inform. Systems Quart.* 42(2):427–464.
- Abrahams AS, Jiao J, Wang GA, Fan W (2012) Vehicle defect discovery from social media. *Decision Support Systems* 54(1):87–97.
- Abrahams AS, Fan W, Wang GA, Zhang ZJ, Jiao J (2015) An integrated text analytic framework for product defect discovery. *Production Oper. Management* 24(6):975–990.
- Abrahams AS, Jiao J, Fan W, Wang GA, Zhang Z (2013) What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems* 55(4): 871–882.
- Adjero D, Beal R, Abbasi A, Zheng W, Abate M, Ross A (2014) Signal fusion for social media analysis of adverse drug events. *IEEE Intelligence Systems* 29(2):74–80.
- Agarwal R, Dhar V (2014) Big data, data science, and analytics: The opportunity and challenge for IS research. *Inform. Systems Res.* 25(3):443–448.
- Agarwal R, Gao G, DesRoches C, Jha AK (2010) The digital transformation of healthcare: Current status and the road ahead. *Inform. Systems Res.* 21(4):796–809.
- Anderson CL, Agarwal R (2011) The digitization of healthcare: Boundary risks, emotion, and consumer willingness to disclose personal health information. *Inform. Systems Res.* 22(3):469–490.
- Aytug H, Koehler GJ (1996) Stopping criteria for finite length genetic algorithms. *INFORMS J. Comput.* 8(2):183–191.
- Bardhan I, Oh J, Zheng Z, Kirksey K (2015) Predictive analytics for readmission of patients with congestive heart failure: Analysis across multiple hospitals. *Inform. Systems Res.* 26(1):19–39.
- Blattberg RC, Byung-Do K, Neslin SA (2008) *Database Marketing: Analyzing and Managing Customers* (Springer, New York).
- Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J. Comput. Sci.* 2(1):1–8.
- Boynton J (2013) How the voice of the people is driving corporate social responsibility. *Harvard Bus. Rev.* (July 17), <https://hbr.org/2013/07/how-the-voice-of-the-people-is>.
- Broniatowski D, Paul MJ, Dredze M (2014) National influenza surveillance through Twitter. *PLoS One* 8(12):e83672.
- Browne J, Manning H, O'Connor C (2015) How to use text analytics in your VoC program. Report, Forrester Research, Cambridge, MA.
- Brynjolfsson E, Geva T, Reichman S (2016) Crowd-squared: Amplifying the predictive power of search trend data. *Management Inform. Systems Quart.* 40(4):941–961.
- Cassino D (2016) The 'wisdom of the crowd' has a pretty bad track record at predicting jobs reports. *Harvard Bus. Rev.* (July 8), <https://hbr.org/2016/07/the-wisdom-of-the-crowd-has-a-pretty-bad-track-record-at-predicting-jobs-reports>.
- Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. *Management Inform. Systems Quart.* 36(4):1165–1188.
- Chen Y, Ganesan S, Liu Y (2009) Does a firm's product-recall strategy affect its financial value? An examination of strategic alternatives during product-harm crises. *J. Marketing* 73(6):214–226.
- Colella-Walsh S (2019) Pradaxa (Dabigatran) trials continue nationwide. Accessed February 10, 2019, <https://www.natlawreview.com/article/pradaxa-dabigatran-trials-continue-nationwide>.
- Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Sci.* 53(9):1375–1388.
- Davies J (2015) 15 voice-of-the-customer best practices linked to organizational maturity. Report, Gartner, Stamford, CT.
- Davies J (2016) How to start creating a voice-of-the-customer strategy. Report, Gartner, Stamford, CT.
- DuMouchel W (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Statist.* 53(3):177–190.
- Esuli A, Sebastiani F (2006) Sentiwordnet: A publicly available lexical resource for opinion mining. *Proc. Language Resources Evaluation Conf.* (European Language Resources Association, Luxembourg), 417–422.
- Fang X, Hu PJH, Li Z, Tsai W (2013) Predicting adoption probabilities in social networks. *Inform. Systems Res.* 24(1):128–145.
- Fenwick N, Leaver S, Kark K, Paderni LS, Blackburn L (2011) Social business strategy: An IT execution plan. Report, Forrester Research, Cambridge, MA.

- Fichman RG, Kohli R, Krishnan R (2011) The role of information systems in healthcare: Current research and future trends. *Inform. Systems Res.* 22(3):419–428.
- Forster AJ, Jennings A, Chow C, Leeder C, van Walraven C (2012) A systematic review to evaluate the accuracy of electronic adverse drug event detection. *J. Amer. Medical Informatics Assoc.* 19(1):31–38.
- Fredericks B (2014) Toyota to pay \$1.2B settlement in vehicle acceleration lawsuit. *New York Post* (March 19), <https://nypost.com/2014/03/19/toyota-to-pay-1-2b-settlement-in-vehicle-acceleration-lawsuit/>.
- Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. *Management Inform. Systems Quart.* 37(2):337–355.
- Hadzi-Puric J, Grmusa J (2012) Automatic drug adverse reaction discovery from parenting websites using disproportionality methods. *IEEE Internat. Conf. Adv. Soc. Networks Anal. Mining* (IEEE, Piscataway, NJ), 792–797.
- Hassan A, Abbasi A, Zeng D (2013) Twitter sentiment analysis: A bootstrap ensemble framework. *IEEE Internat. Conf. Soc. Comput.* (IEEE, Piscataway, NJ), 357–364.
- Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. *Management Inform. Systems Quart.* 28(1):75–105.
- Hora M, Bapuji H, Roth AV (2011) Safety hazard and time to recall: The role of recall strategy, product defect type, and supply chain player. *J. Oper. Management* 29(7):766–777.
- Jin HW, Chen J, He H, Kelman C, McAullay D, O’Keefe CM (2010) Signaling potential adverse drug reactions from administrative health databases. *IEEE Trans. Knowledge Data Engrg.* 22(6):839–853.
- Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C (2015) Text and data mining techniques in adverse drug reaction detection. *ACM Comput. Surveys* 47(4):Article 56.
- Kitchens B, Dobolyi D, Li J, Abbasi A (2018) Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *J. Management Inform. Systems* 35(2):540–574.
- Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? *Proc. 19th ACM Internat. Conf. World Wide Web* (ACM, New York), 591–600.
- Lardon J, Abdellaoui R, Bellet F, Asfari H, Bousquet C (2015) Adverse drug reaction identification and extraction in social media: A scoping review. *J. Medical Internet Res.* 17(7):1–16.
- Lau R, Liao S, Wong KF, Dickson K (2012) Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. *Management Inform. Systems Quart.* 36(4):1239–1268.
- Lazer D, King G, Vespignani A (2014) The parable of Google Flu: Traps in big data analysis. *Science* 343(6176):1203–1205.
- Lindquist M (2008) VigiBase, the WHO Global ICSR Database System. *Drug Inform. J.* 42(5):409–419.
- Liu M, Hinz ERM, Xu H (2013) Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *J. Amer. Medical Inform. Assoc.* 20(3):420–426.
- Liu X, Chen H (2013) AZDrugMiner: An information extraction system for mining patient-reported adverse drug events in online patient forums. Zeng D, Yang CC, Tseng VS, Xing C, Chen H, Wang F-Y, Zheng X, eds. *Proc. Internat. Conf. Smart Health* (Springer, Berlin), 134–150.
- Provost F, Fawcett T (2013) *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O’Reilly Media, Sebastopol, CA).
- Provost F, Martens D, Murray A (2015) Finding similar mobile consumers with a privacy-friendly geosocial design. *Inform. Systems Res.* 26(2):243–265.
- Ritter JM (2008) Minimising harm: Human variation and ADRs. *Brit. J. Clinical Pharmacol.* 65(4):451–452.
- Sampathkumar H, Chen XW, Luo B (2014) Mining adverse drug reactions from online healthcare forums using hidden Markov model. *BMC Medical Informatics Decision Making* 14(1):Article 91.
- Sarker A, Gonzalez G (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomedical Inform.* 53(February):196–207.
- Sarker A, Ginn R, Nikfarjam A, O’Connor K, Smith K, Jayaraman S, Gonzalez G (2015) Utilizing social media data for pharmacovigilance: A review. *J. Biomedical Inform.* 54(April):202–212.
- Schmidt-Subramanian M, Manning H, Czarnecki D (2014) The state of voice-of-the-customer programs: It’s time to act. Report, Forrester Research, Cambridge, MA.
- Schweinsberg C (2012) Toyota agrees to \$1.1 billion settlement in unintended acceleration lawsuit. Accessed May 1, 2017, <https://www.wardsauto.com/industry/toyota-agrees-11-billion-settlement-unintended-acceleration-lawsuit>.
- Sharif H, Zaffar F, Abbasi A, Zimbra D (2014) Detecting adverse drug reactions using a sentiment classification framework. *Proc. 6th ASE Internat. Conf. Soc. Comput.* (IEEE, Piscataway, NJ), 231–240.
- Shmueli G, Koppius OR (2011) Predictive analytics in information systems research. *Management Inform. Systems Quart.* 35(3):553–572.
- Song Y, Sahoo N, Srinivasan S, Dellarocas C (2014) Uncovering path-to-purchase segments in large consumer population. *24th Workshop on Information Technologies and Systems, Auckland, NZ.*
- Sunstein C (2006) When crowds aren’t wise. *Harvard Bus. Rev.* 9(September):20–21.
- Surowiecki J (2005) *The Wisdom of Crowds* (Anchor Press, New York).
- Thomas K (2014) \$650 Million to settle blood thinner lawsuits. *New York Times* (May 28), <https://www.nytimes.com/2014/05/29/business/international/german-drug-company-to-pay-650-million-to-settle-blood-thinner-lawsuits.html>.
- Turney PD, Littman ML (2003) Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inform. Systems* 21(4):315–346.
- Wasko MM, Faraj S (2005) Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *Management Inform. Systems Quart.* 29(1):35–57.
- White RW, Tatonetti N, Shah N, Altman RB, Horvitz E (2013) Web-scale pharmacovigilance: Listening to signals from the crowd. *J. Amer. Med. Inform. Assoc.* 20(3):404–408.
- Xu R, Wang Q (2014) Large-scale combining signals from both biomedical literature and FAERS to improve post-marketing drug safety signal detection. *BMC Bioinformatics* 15(1): Article 17.
- Yang CC, Yang H, Jiang L (2014) Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. *ACM Trans. Management Inform. Systems* 5(1):Article 2.
- Yang M, Wang X, Kiang MY (2013) Identification of consumer adverse drug reaction messages on social media. *Proc. Pacific-Asia Conf. Inform. Systems* (Association for Information Systems, Atlanta), Paper 193.
- Zabin J, Nail J, Wilder SK (2011) Gleansight social intelligence. Report, Gleanster Research, Pleasanton, CA.
- Zeng D, Chen H, Li S (2010) Social media analytics and intelligence. *IEEE Intelligence Systems* 25(6):13–16.
- Zimbra D, Abbasi A, Zeng D, Chen H (2018) The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Trans. Management Inform. Systems* 9(2):Article 5.