Forthcoming in MIS Quarterly

# Timely, Granular, and Actionable: Designing a Social Listening Platform for Public Health 3.0

**Brent Kitchens**
IT & Innovation Area
McIntire School of Commerce
University of Virginia
bmk2a@virginia.edu

**Jennifer Claggett**
MIS Area
School of Business
Wake Forest University
claggejl@wfu.edu

**Ahmed Abbasi**
Human-centered Analytics Lab
Department of IT, Analytics, & Operations
University of Notre Dame
aabbasi@nd.edu

## Abstract

Every day patients access and generate online health content through a variety of online channels, creating an ever-expanding sea of data in the form of digital communications. At the same time, proponents of public health have recently called for timely, granular, and actionable data to address a range of public health issues, stressing the need for social listening platforms that can identify and compile this valuable data. Yet previous attempts at social listening in healthcare have yielded mixed results, largely because they have failed to incorporate sufficient context to understand the communications they seek to analyze. Guided by Activity Theory to design HealthSense, we propose a platform for efficiently sensing and gathering data across the web for real time analysis to support public health outcomes. HealthSense couples theory-guided content analysis and graph propagation with graph neural networks (GNNs) to assess the relevance and credibility of information, as well as intelligently navigate the complex online channel landscape, leading to significant improvements over existing social listening tools. We demonstrate the value of our artifact in gathering information to support two important exemplar public health tasks: 1) performing post market drug surveillance for adverse reactions and 2) addressing the opioid crisis by monitoring for potent synthetic opioids released into communities. Our results across data, user, and event experiments show that effective design artifacts can enable better outcomes across both automated and human decision-making contexts, making social listening for public health possible, practical, and valuable. Through our design process, we extend Activity Theory to address the complexities of modern online communication platforms, where information resides not only within the collection of individual communication activities, but in the complex network of interactions between them.

**Keywords:** public health 3.0, social listening, activity theory, online platforms, computational design, graph neural networks, opioid crisis, pharmacovigilance

## 1 Introduction

Users access and generate health-related information through a variety of online platforms, creating an ever-expanding sea of data. These interactions range from online health communities (Yan and Tan 2014) to patient portals (Peacock et al. 2017), general microblogs such as Twitter (Barnes et al. 2019), and open discussion boards such as Reddit (Park et al. 2018). Prior research has focused on how patient interactions with online resources may benefit individual health outcomes, such as managing chronic disease (Liu et al. 2020), patient education (Hansen 2008), patient emotional support (Yan and Tan 2014), and clinical decision support (Wright et al. 2009).

But beyond individual benefits, there is significant, untapped potential in leveraging these vast and various digital communications to address greater public health issues (Fichman et al. 2011), such as combating the opioid crisis (Bowen et al. 2019), detecting adverse drug events (Adjeroh et al. 2014), understanding e-cigarette trends (Cole-Lewis et al. 2016), and tracking disease prevalence (Yang et al. 2013). The critical and oft-ignored first step in accomplishing these goals is the employment of effective social listening platforms to efficiently gather relevant information to support time-sensitive analytics. Indeed, a movement toward "public health 3.0" has made specific calls to address the dearth of "timely, granular, and actionable" data to support public health initiatives (Wang and DeSalvo 2018), and data from social platforms to support public health research and practice (Pagoto et al. 2019). Social listening platforms have been used to great effect in other areas – for instance, by marketers to understand customer opinions (Davis and Logan 2019; Hewett et al. 2016). However, significant limitations in available social listening platforms hinder them from effectively supporting public health informatics.

The recent COVID-19 epidemic has put a spotlight on public health issues, and the need for data to support time-sensitive decisions. For instance, Apple and Google partnered to provide health authorities with anonymized public movement data and contact-tracing capabilities (Apple Media 2020). However, social listening initiatives at public health agencies remain limited, with significant opportunity for development. Per conversations with US Health and Human Services officials in July 2021, aside from the public movement data, HHS has no partnerships or initiatives to utilize online platform data. The CDC has issued guidance for usage of "social listening and monitoring tools," but only lists commercial platforms with limited data sources (CDC 2021) This leaves significant room for improvement in utilizing such data to support public health analytics.

Social listening platforms applied to the healthcare context represent extreme cases regarding both the opportunity for improving public health and the complexity involved in the gathering, parsing, and understanding of relevant data (Boudry 2015; Pour and Jafari 2018). Previous attempts at social listening platforms in healthcare have had mixed results. In a famous example, Google Flu Trends (GFT) was designed to predict influenza cases using search data. Despite initial promise, the project failed (Lazer et al. 2014). Researchers pointed out that the narrow view taken by GFT ignores issues such as context, credibility, and the inherently omni-channel nature of online interactions (Broniatowski et al. 2014; Lazer et al. 2014). Traditionally, social listening has been considered a technology problem, but solely technology-centered solutions like the original GFT algorithm often fall short. It is challenging to extract meaning from online platform data, as communication is a human process (Abbasi et al. 2018). Effective health-focused social listening platforms must be designed from a socio-technical perspective.

In order to effectively leverage social listening to address time-sensitive public health outcomes, information must be gathered efficiently from a wide variety of sources spread across the digital landscape. Currently available platforms tend either to focus on a highly curated set of sources (Davis and Logan 2019; Sarker et al. 2015), or slowly crawl the web to collect data-warehouse-like snapshots of data, outdated before it is captured (Kumar et al. 2017). To design an improved social listening platform capable of supporting public health use cases, we utilize and extend Activity Theory (Chen et al. 2013; Engeström 1987; Valecha et al. 2019) to imbue it with the context of digital communications, allowing it to efficiently identify, cull, and gather the most relevant information across a variety of online platforms for use in time-sensitive analytics. Activity Theory has been utilized to great success in understanding individual communication processes. However the phenomenon of modern communication through myriad interconnected online platforms represents significantly more than the sum of these individual communications. To truly capture meaning from online discourse, this theory must be expanded to consider how information arises from the complex network of interactions between individual content, authors, channels, communities, and platforms. As we demonstrate, it is not only within these communication activities that information resides, but also (and often more importantly) between.

Following a computational design research approach (Rai 2017; Padmanabhan et al. 2022), we propose HealthSense, a design artifact for social listening to collect timely, granular, actionable data in support of public health 3.0 analytics. The design relies critically on our extension of Activity Theory to incorporate characteristics of the *multiplex relationships* between authors, channels, communities, and content which arise as information propagates through online platforms. The context concealed in these relationships provides the key to designing an effective social listening system. To demonstrate the efficacy and value of HealthSense, we run a series of experiments on a large health dataset encompassing 37 million data points related to opioids and adverse drug events. Data experiments reveal that HealthSense efficiently identifies over 90% of task-relevant content from only 20% of the data – much faster than comparison methods including state-of-the-art deep learning techniques. User and event detection experiments with a major pharmaceutical drug safety team show that HealthSense's performance facilitates better automated detection and allows analysts to make more accurate decisions related to adverse drug events.

Our work makes several contributions to research and practice. First, we develop a novel artifact guided by Activity Theory that seamlessly combines relevance, credibility, and cross-channel landscape assessment to enable markedly better and more timely public health listening capabilities. Second, through our design process, we propose an extension to Activity Theory to include multiplex relationships between activities that are critically important, particularly when analyzing communication activities within highly connected online platforms. Third, we extend knowledge of graph neural networks through two novel extensions which improve the performance of these methods. Fourth, more broadly, we show that in time-sensitive environments at the intersection of Big Data and machine learning, artifacts guided by human-centered theories and intuition are critical complements to automated AI-driven techniques. Finally, our rigorous evaluation across data, event, and user experiments show that effective design artifacts can make social listening for public health possible, practical, and valuable.

# 2 Background

## 2.1 The State of Social Listening Research

The use of data from online platforms to achieve organizational or societal goals has been explored in many use cases. Among the most prevalent has been opinion mining – analyzing user generated opinions shared on social media platforms regarding brands and products for various marketing purposes (Adamopoulos et al. 2018). Content from health forums and other online platforms has been analyzed to support health outcomes, such as the design of tools for detecting adverse drug events (Adjeroh et al. 2014) and for understanding how patient support forums assist in the management of chronic diseases (Dadgar and Joshi 2018). However, a common characteristic among the vast majority of studies that seek to leverage data from online platforms is that they are largely focused on explaining a phenomenon or providing a proof of concept for their methods, and take data acquisition for granted – often relying on pre-collected (possibly proprietary) data or a convenience sample from siloed data sources, such as the Twitter API or social monitoring tools such as Brandwatch (Davis and Logan 2019; McClellan et al. 2017). As discussed regarding Google Flu Trends, using a single, narrow data source without context can be highly problematic.

Web crawlers or spiders provide a method for broader acquisition of large volumes and varieties of online content (Kobayashi and Takeda 2000). Traditionally a set of seed URLS are provided to a crawler, which recursively traverses out-links from those URLS and the pages they point to, eventually collecting millions of pages (or more) for analysis (Cho et al. 1998). While these tools are effective, they are indiscriminate in the content they collect, and very slow. Focused crawlers attempt to address these issues by seeking out content based on specific, predefined requirements (Chen et al. 2003). They tend to manage their crawl frontier (i.e., the pages to be collected) by topic prioritization (Pant and Srinivasan 2005). That is, the crawler typically prioritizes and determines whether to collect a given URL based on lexical analysis to determine topic relevance of content surrounding the link in the source page and/or in ancestor pages already collected (Aggarwal et al. 2001; Diligenti et al. 2000; Fu et al. 2012; Pant and Srinivasan 2005).

While focused crawlers do provide more topic-relevant content, most do not provide the efficiency needed to support time-sensitive analysis. More importantly, many of these crawlers only consider lexical features to determine topic relevance. Some add features such as graph-based analysis (Chau and Chen 2007) and sentiment in addition to topic (Fu et al. 2012). Others have explored crawling web and social media for event-related information (Farag et al. 2018). However, for the time-sensitive analysis required for many public health use cases, there is need for a social listening platform that can holistically, yet efficiently, consider the context and complexities of human communication from a broad spectrum of online resources.

## 2.2 Node Embeddings and Graph Neural Networks

Advances in neural networks have introduced new ways to represent graph-based data. Node embeddings allow local network neighborhood information for each node in the network to be encoded into a fixed-length vector representation. For instance, techniques such as DeepWalk (Perozzi et al. 2014) and Node2Vec (Grover and Leskovec 2016) embody similar design intuitions to Word2Vec style word embeddings – that is, graph random walks are run through a single-layer neural network. More recent work on graph neural networks (GNN) allows node embeddings to be learned via message-passing as part of an end-to-end learning framework, based on how effectively they can support downstream classification tasks (e.g., node, edge, graph, etc.). Examples include graph convolutional networks (Kipf and Welling 2017) and GraphSage (Hamilton et al. 2017). Graph attention networks extend these methods by replacing fixed edge weights with parametric attention weights (Velickovic et al. 2018). A historical limitation of GNNs has been the assumption of homogeneous node and edge types (Wu et al. 2021). Heterogeneous graph neural networks have been proposed as a way to allow inclusion of different explicitly defined node types (Zhang et al. 2019), as well as attention mechanisms (Wang et al. 2019).

For social listening platforms, GNNs afford opportunities for incorporating rich, domain-adapted graph representations. This is part of an exciting trend where advancements in machine learning enable the development of artifacts that better capture the richness of key social-technical phenomena, allowing closer alignment with underlying theories (Rai 2017; Padmanabhan et al. 2022; Yang et al. 2022). We propose a novel embedding fused with graph propagation to capture complex cross-channel individual and network-level activities. By overcoming current limitations of GNNs in the context of social listening such as sparse labels, scalability to larger graphs, and effectively considering different types of heterogeneity (Wu et al. 2021; Dai et al. 2022; Bojchevski et al. 2020), our artifact outperforms existing focused crawling and GNN methods.

## 3 An Activity Theoretic Perspective

Digital communication is a human process, and therefore may only be truly understood from a socio-technical vantage point, with appropriate consideration for context that supports sense-making (Abbasi et al. 2018). Understanding this complex digital communication process is key in the design of an efficient social listening platform capable of supporting timely and relevant insights. With this goal in mind, we inform our social listening platform design by (1) utilizing Activity Theory to examine the creation of online content and (2) extending Activity Theory to frame digital communication as a *multiplex communication activity*. Activity Theory recognizes that an activity can be thought of as an interaction between a *subject* (usually a human being) and an *object*, in order to generate an *outcome* (Leontev 1978), often using a *tool* (Kaptelinin and Nardi 2006). Engeström (1987) expanded the theory to recognize that activities are carried out within the context of a *community*. Activity Theory is not a predictive theory, but a framework for orienting an observer to better understand a complex, real-life problem (Kuutti 1991). It has been used often in studying human-computer interaction, as the use of technology is dependent on the complex, dynamic social environments where it takes place (Allen et al. 2013), and to inform the design of information systems (Chen et al. 2008, 2013; Korpela et al. 2002).

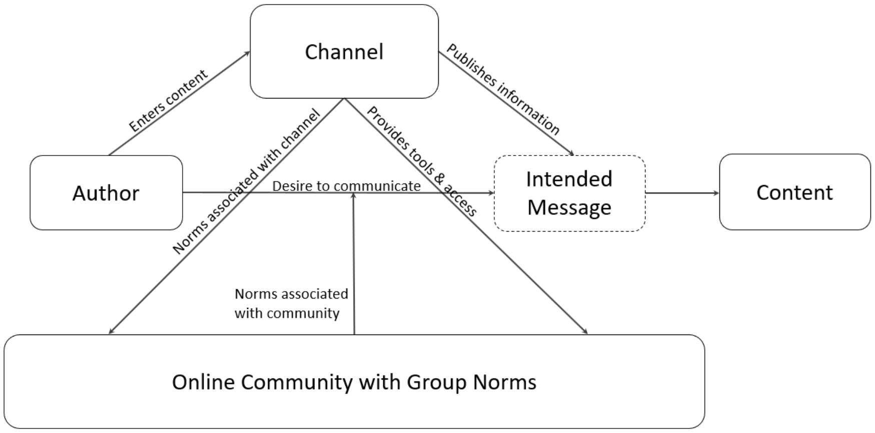
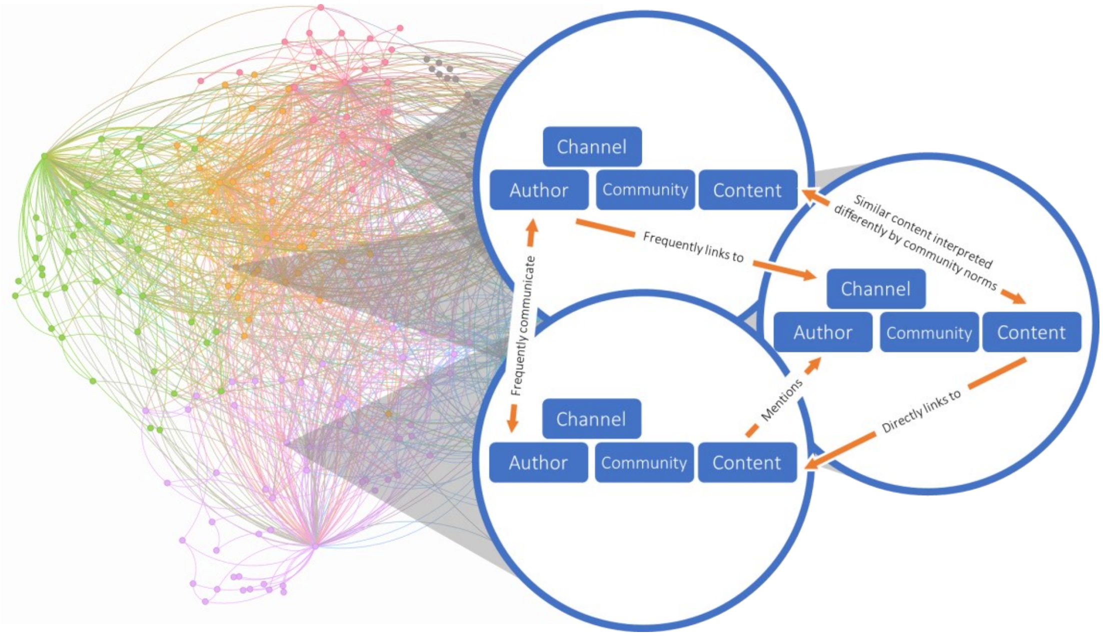


**Figure 1**: Activity framework showing Online Context Examples

Designed as a descriptive meta-theory or framework, Activity Theory should be interpreted for the context of interest (Chen et al. 2013; Engeström 1987), which for our purposes is the creation of online health-related content. As depicted in Figure 1, in our context we understand the subject to be a user posting content, and the online channel (e.g., Twitter, web pages, blogs, etc.) to be the tool. Further, online communities provide important context, including group norms, formal and informal rules, acronyms, and coded language with special meaning.

While this view is informative in understanding the activities of individuals, our goal is to leverage large volumes of online content created by the combination of these activities in order to understand public health phenomenon. As shown in Figure 2, information propagates within and across online platforms through communication activities that are not only highly interconnected, but connected by multiplex relationships of their constituent components. That is, not only may pieces of content be related, but the author of one piece of content may be related to a second community where other content was created; or the authors of two unrelated pieces of content may regularly correspond in other ways on the platform; or similar content may be shared across two separate channels. The interactions between the individual elements of communication activities (importantly author, channel, community, and content) give rise to a multiplex network of inter-activity relationships which can provide rich information to social listening platforms.

While research on Activity Theory has noted the potential to consider the interactions between activity processes, this has largely been limited to examining pairs of activity processes to understand co-production of outputs (Allen et al. 2013; Chen et al. 2013; Effah and Adam 2021). Outside of Activity Theory, other research streams have recently emphasized the importance of inter-connected activities on online platforms. Most notably, there is a quickly growing literature regarding omnichannel user behavior, particularly as it relates to designing and measuring marketing interventions and outcomes (e.g. Cui et al. 2021; Sun et al. 2022). However, even this emergent research focuses only on a limited subset of the multiplex relationships among activities on online platforms (namely, how individual users interact with multiple channels). We propose that incorporation of a full complement of these multiplex inter-activity relationships is crucial to extracting information from the cacophony of communication activities on online platforms. When designing artifacts, kernel theories from the natural or social sciences are often used to guide meta-requirements and meta-design (Walls et al. 1992). However, existing individual theories are rarely used without adaptation, and design often requires integrating multiple theoretical perspectives (Arazy et al. 2010). This extension to Activity Theory represents a core component informing our meta-design, which we demonstrate to be critical in our artifact instantiation. Moreover, in order to meet social listening needs in the public health 3.0 context, we couple our extended Activity Theoretic perspective with characteristics from the context-based information quality literature.



Note: The relationships noted in this stylized figure are only exemplar as the full complement of exponential relationships would be difficult to depict—each component of each activity has a potential relationship with each component of all other activities

**Figure 2**: Multiplex relationships among communication activities on online platforms

# 4 Requirements & Design

The public health domain is undergoing a paradigm shift into what has been termed "public health 3.0" (DeSalvo et al. 2017). A core mantra of this movement is the call for "timely, granular, actionable data" to support public health informatics (Wang and DeSalvo 2018). More generally, these calls point to the need for data that is relevant and useful to the task of public health management, which aligns with the context-based based definition of information quality (Nelson et al. 2005). Context-based information quality considers output from a system and determines if it is helpful towards the task at hand by identifying and measuring the output across relevant dimensions. As is often noted, there is no generally agreed-upon definition of information quality, and the literature acknowledges it is best considered a multidimensional constructure that has both objective and subjective components depending on the context (Arazy and Kopak 2011). Table 1 represents a summary of the significant research that has been performed to conceptualize contextual information quality in various contexts, resulting in identification of a wide range of relevant dimensions. In analyzing these dimensions, we identify four high-level themes which inform our requirements: Timeliness, Relevance, Credibility, and Completeness. In the following sections, we outline each of these requirements as well as how our extension of Activity Theory informs the design of our artifact to accomplish them, as summarized in Table 2.

**Table 1: Conceptualization of context-based information quality**

| Source | Timeliness | Relevance | Credibility | Completeness | Other * |
|---|---|---|---|---|---|
| (Strong et al. 1997) | Timeliness | Relevancy, Value-Added | Accuracy, Believability, Consistency, Objectivity | Completeness, Amount of data | Accessibility, Interpretability, Ease of Understanding, Conciseness |
| (Nelson et al. 2005) | Currency | Meaningfulness (as part of Accuracy) | Believable,<br>Consistent, Correct (as part of Accuracy) | Completeness | Format |
| (Lee et al. 2002) | Currency, Cycle time, Non-volatility, Timeliness | Content, Essentialness, Informativeness, Importance, Meaningfulness, Precision, Relevance, Usefulness, Value-added | Accuracy, Believability, Consistency, Correctness, Credibility, Factual, Free from bias, Objectivity, Reliability, Reputation, Validity | Appropriate amount, Attribute granularity, Completeness, Comprehensiveness, Level of detail, Quantity, Sufficiency | Unambiguous, Usage |
| (Knight and Burn 2005) | Timeliness, Efficiency | Relevancy, Useability, Useful, Value-added | Accuracy, Believability, Consistency, Objectivity, Reliability, Reputation | Amount of data, Completeness | Accessibility, Availability, Concise, Navigation, Security, Understandability |
| (Ge et al. 2011) | Timeliness | Relevancy, Interpretability, Value-added | Accuracy, Believability, Consistency, Objectivity, Reliability, Reputation | Appropriate amount, Completeness | Accessibility, Ease of manipulation, Ease of representation/ understanding, Security |
| (Arazy et al. 2017; Arazy and Kopak 2011) | Accessibility (discussed only and includes timeliness) | | Accuracy, Objectivity | Completeness | Representation |
| (Zheng et al. 2013) | Timeliness | Value-added | Objectivity, Reliability | Richness | Format |
| (Todoran et al. 2015) | Timeliness | Relevancy, Interpretability | Accuracy, Consistency, Correctness, Integrity, Objectivity Reliability, Reputation | Completeness, Data amount | Accessibility, Security, Understandability |

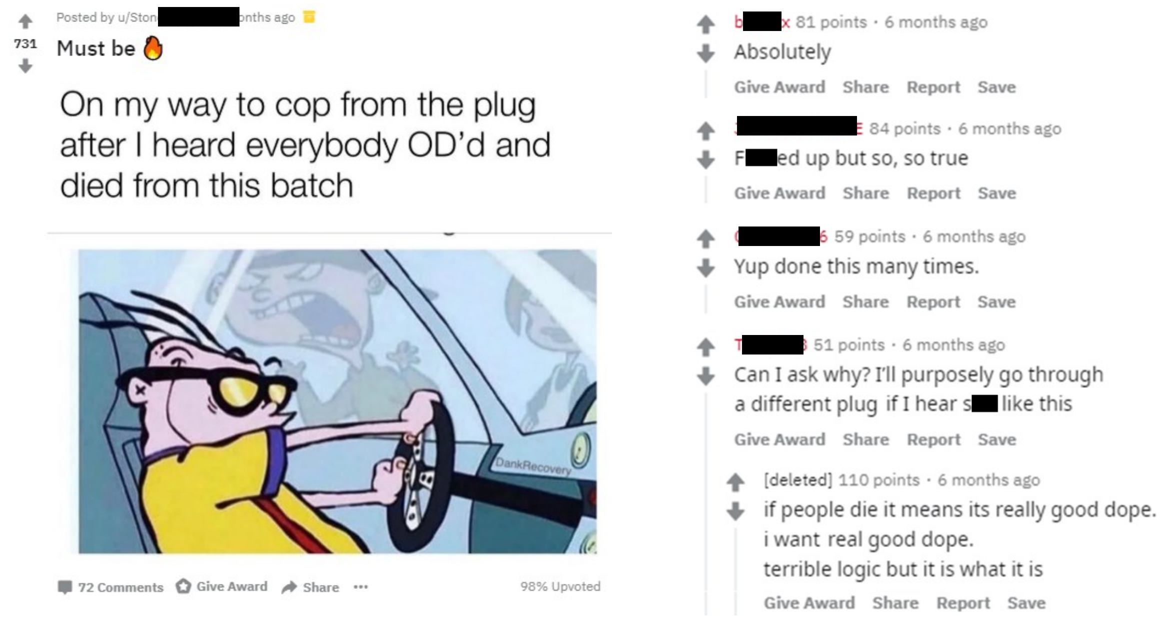
**Table 2:** Requirements and design elements

| Activity Theoretic Meta-Requirements | | Information Quality Meta-Requirements | | | |
|---|---|---|---|---|---|
| | | **0. Gather *Timely* data from a variety of online platforms to support time-sensitive analysis** | | | |
| | | **1. Evaluate *Relevance*** | **2. Evaluate *Credibility*** | **3. Navigate Channel Landscape for *Completeness*** | |
| **Individual communication activities** | **Author *tendencies*** | Learn author tendencies from aggregate collection of authored documents (a) | | | **Design Elements** |
| | **Channel *context*** | Leverage channel-specific patterns (e.g., web, blog, forum, and social media) (b) | | Incorporate channel source context to predict relevance of a given path (c) | |
| | **Community *norms*** | Capture group-specific linguistic norms through use of lexicons, libraries, and dictionaries (d) | Incorporate group norms into seeding (e) | | |
| | **Content *language*** | Analyze linguistic features to determine topic and sentiment (f) | | Incorporate content topic and sentiment information into graphs (g) | |
| **Communication network** | **Inter-activity *relationships*** | Incorporate author and site relationships through features derived from multi-level graph (h) | Analyze multi-level bi-directional author, site and document-level graphs representing inter-activity relationships (i) | | |
| | **Information *propagation*** | Incorporate spatial dynamics through graph propagation and convolution (j) | | | |

(a)-(j) During our analysis we run an ablation analysis to demonstrate the importance of each lettered element in our artifact performance. Please see Tables 7 and 8 for details.

## 4.1 Motivating Case: Opioid Epidemic

Although our design is generalizable to a large variety of public health issues, it is useful to consider a motivating case to describe the requirements for a social listening platform for public health. The opioid epidemic is a major public health crisis, particularly in developed countries such as the United States (Schwetz et al. 2019). Each day 128 people die of overdoses from opioids (Wilson 2020), i.e., drugs naturally derived from poppy plants and their synthetic counterparts. In the past several years, there has been a drastic increase in overdoses from synthetic opioids such as Fentanyl and Carfentanil (an elephant tranquilizer and chemical weapon), which are hundreds to thousands of times as potent as heroin. Drug dealers often cut heroin with these because they are cheaper and more potent, creating uncharacteristically strong batches of difficult to dose drugs – leading to sudden spikes in overdoses (Shoff et al. 2017).



Posted by u/Ston[redacted]nths ago

731

Must be 🔥

On my way to cop from the plug after I heard everybody OD'd and died from this batch

72 Comments · 98% Upvoted

b[redacted]x 81 points · 6 months ago

Absolutely

[redacted]E 84 points · 6 months ago

F[redacted]ed up but so, so true

[redacted]5 59 points · 6 months ago

Yup done this many times.

T[redacted]3 51 points · 6 months ago

Can I ask why? I'll purposely go through a different plug if I hear s[redacted] like this

[deleted] 110 points · 6 months ago

if people die it means its really good dope.
i want real good dope.
terrible logic but it is what it is

Note: In these social networks, "cop" is a term for purchasing drugs, and "plug" is used to describe a prolific or notorious drug dealer. So, the meme depicts a person in a hurry to jump in their car and buy from a batch because overdoses and deaths were reported. Many users immediately posted agreement with the sentiment.

**Figure 3**: Reddit posting (and related comments) making a joke about seeking out a batch of drugs that caused overdoses and deaths due to its high potency

Despite the often fatal outcomes, drug dealers are incentivized to make these dangerous batches because they are highly sought after by customers, as Joseph Pinjuh, chief of the narcotics unit for the U.S. attorney in Cleveland explained: "[Drug users] know that's the high that'll take you right up to the edge, maybe kill you, maybe not…That's the high that they want." (AP 2016). As news spreads over social media of users overdosing, there is a surge of interest in that batch of drugs (Overbeek and Janke 2018). Figure 3 presents a Reddit post illustrating this mindset.

When these potent batches arrive in an area, word spreads rapidly and the situation can escalate quickly. In 2016 Akron police reported 25 overdoses in a three day period and near-by Columbus reported 10 overdoses in a nine-hour period (DeMio 2016). This phenomenon provides a clear example where timely detection of the situation (i.e., a new batch of especially strong opioids has arrived in a certain local) could result in better health outcomes. Normally, hospitals and EMS vehicles have a limited supply of Naloxone (common trade name, Narcan) which is an injectable medicine used to reverse the effects of opioids in an emergency situation.

We interviewed the program director of a local clinic specialized in treating opioid addicts and president of the local volunteer rescue squad. She explained that a "good" batch (meaning cut with synthetic opioids, such that it is dangerous and especially potent) often causes a spike in overdoses for 24 to 72 hours after its release. In this case, first responders may easily find themselves in situations with inadequate supplies of Naloxone on hand, with multiple overdose victims present and/or individual victims requiring multiple doses. If first responders were aware of the likely high rate of these heavy overdoses in advance, they could stock more Naloxone in their vehicles. An interviewed emergency room physician[1] agreed and further explained that if he were made aware that a strong batch of Fentanyl- or Carfentanil- laced opioids had been released in the area, he would treat patients presenting with overdose symptoms more aggressively in the emergency room. These sentiments echo research which has suggested that better data could have significant impacts in combatting opioid overdoses (Bowen et al. 2019; Saloner et al. 2020).

Social media, websites, blogs, and forums are used by drug dealers and drug seekers to spread this information about batch strength and availability within the opioid using community (Overbeek and Janke 2018; Pandrekar et al. 2018). We propose that a social listening platform can provide the timely, granular, actionable information needed to prepare local responders for these strong batches of opioids. In Table 2, we outline how Activity Theory intersects with the requirements of a social listening platform for public health to inform the design of such a system. In the remainder of this section, we will describe these requirements and illustrate selected design elements within the context of this use case of surveillance to detect dangerous batches of opioids. To provide concrete examples in line with our motivating case, we collected data from Reddit groups r/opiates and r/opiaterollcall. Although we use this example to illustrate the key system requirements and underlying gaps they address, it is important to note that our work is relevant to social listening for various types of health events. We later evaluate our proposed artifact on multiple health use cases including opioids and adverse drug events.

### 4.2 Meta-Requirement 0: Timeliness

Timeliness is the primary requirement which underlies the motivation for our design, in that many of the other dimensions are easily addressable without this constraint. Capabilities for timely use of data (often referred to as "velocity") have been a primary driver for deriving value from the "big data" movement in recent years (Kitchin and McArdle 2016). From a contextual quality standpoint, timeliness or currency relates to whether data is sufficiently "up to date" for use in a given task (Nelson et al. 2005, Strong et al. 1997). Within the health context, timeliness is driven by the urgency of many public health issues. A potent batch of opioids may cause dozens of overdoses within hours or days without intervention (DeMio 2016; Shoff et al. 2017). Failure to identify adverse drug events in a timely manner may cost pharmaceutical companies significant sums and leave masses of patients suffering or worse (Adjeroh et al. 2014). Disease epidemics can explode overnight (Yang et al. 2013). Furthermore, as the overall available digital content and noise grows at a greater pace than the sub-set of relevant, time-critical content, the "need for speed" cannot be overstated (Boldi et al. 2018). The proponents of public health 3.0 also note the necessity of removing silos around data and achieving wide interoperability (Wang and DeSalvo 2018). Indeed, identification of timely, granular, actionable data might be simple if it were available in a single or even small set of sources with open APIs. Unfortunately, content providers, patients, and other users post relevant information across a vast and varied online landscape (Boudry 2015): from social media outlets (Twitter, Reddit); to websites (WebMD.com, Medscape.com), to patient forums (HealthBoards.com, MedHelp.com), to blogs (MothersInMedicine.com, AmbulanceDriverFiles.com), and beyond. The integration of insights from across these various sources is critical in creating effective analytics that can impact public health (Lazer et al. 2014).

The opioid case provides further examples of additional considerations for timely identification of information across a variety of sources. As discussed, there is strong agreement among healthcare professionals that the *prompt* detection of strong batches of opioids in the community would save lives and improve medical outcomes. In addition, sites, forums, blogs, and social media groups hosting this illicit information are highly dynamic and ephemeral (Hayes et al. 2018). They tend to pop in and out of existence quickly, often being found by law enforcement groups or site moderators and shifting to new platforms. Even if the web resource continues to exist (e.g., a drug forum), specific pieces of relevant content (e.g., a specific post) may be deleted almost as soon as it is posted. For instance, Reddit has a culture of user-moderation to enforce group norms and remove content that is illegal, illicit, or otherwise contrary to Reddit policies or goals of the community (Squirrell 2019). In historical data examined for our motivating case, 8.1% of posts were deleted or removed (2.2% within one day, 5.3% within 7 days). These posts are often removed precisely because they include information relevant to the social listening objective, such as mentioning locations and/or attempts at sourcing drugs, often specifically focused on potent batches. This further increases the pressure for efficient social listening.

### 4.3 Meta-Requirement 1: Evaluate Relevance

Beyond the overarching requirement of timeliness, the first theme we note within the contextual information quality literature is that of *relevance* to the problem at hand. Relevance refers to the degree of applicability and usefulness of information for a given task (Lee et al. 2002; Todoran et al. 2015), or the extent to which it is beneficial and provides advantage (Zheng et al. 2013). Within the online health information context, there is a variety of information from many sources which must be evaluated for relevance to the social listening task. Because relevance in a social listening context must be determined before retrieving content, existing tools typically evaluate content surrounding the link on ancestor pages (Farag et al. 2018; Fu et al. 2012; Pant and Srinivasan 2005). However, these basic approaches miss a significant amount of information.

For instance, online platforms are often utilized as people research medical information and express their opinions and emotions (Bender et al. 2008). This process creates rich and abundant content using group- or channel- specific semantic communication features such as slang, abbreviations, and emoticons (Smith et al. 2012; Waterloo et al. 2018). Both within channels as well as within communities formed by groups of users, specialized language or content generation patterns are often developed or used, and these must be understood in order to interpret message relevance (Squirrell 2019). For example, in Figure 3, we see unusual word choices that mean something specific to the community such as “cop” referring to purchase and “plug” referring to a drug dealer in this context.

In addition to the content, other aspects of digital communication creation can be used to help fully interpret the relevance of the information. Authors play a variety of social roles within online communities (Benamar et al. 2017), and uncovering these nuances of author relationships within the multiplex of communication activities can help contextualize the posted content and assess relevance. For example, in the Reddit opioid context, some users primarily ask questions and some answer; others seek or offer drugs; some seek or provide emotional support; and there are those who post tips on harm reduction, or copping/using (Overbeek and Janke 2018; Pandrekar et al. 2018). Figure 4 shows a network of all communications on opiate subreddits among a sample of users, illustrating how various users interact over time. Information propagation and relationships within the author graph provides latent information about these author roles that helps contextualize content. For instance, the active addict (red triangle) who posts controversial information about sourcing drugs has frequent run-ins with the moderator (green square). In order to assess the relevance of digital communications, artifact design should consider not only content, but also how information propagates through the multiplex relationships between content, authorship, channel, and community aspects, all of which provide valuable context.

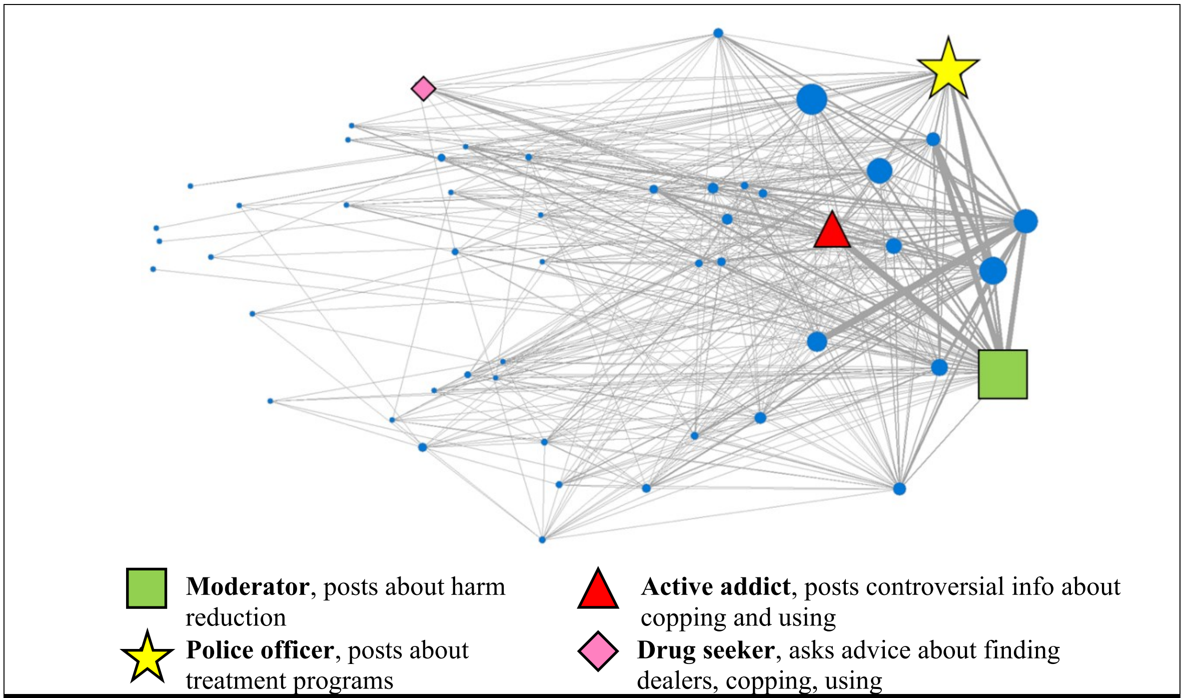

🟩 **Moderator**, posts about harm reduction
🔺 **Active addict**, posts controversial info about copping and using
⭐ **Police officer**, posts about treatment programs
◆ **Drug seeker**, asks advice about finding dealers, copping, using

**Figure 4**: All communications across opiate subreddits among a sample of users from 7/1/13 to 6/30/17. Size of arrow represents frequency of comments of one user in response to another. Selected users are characterized based on their observed posting behaviors.

### 4.4 Meta-Requirement 2: Evaluate Credibility

The next theme of contextual information quality that applies to the design of social listening tools is credibility. To be useful, information must be believable, free-of-error, and correct (Lee et al. 2002). Information quality research often considers accuracy, which is an intrinsic trait (Arazy and Kopak 2011), yet the perception of accuracy is involved in credibility assessments (Fogg 2003). Further, the reputation of the source must also be taken into consideration (Todoran et al. 2015; Zheng et al. 2013). Fake websites, spam, and other non-credible information are especially troubling in the healthcare context, and the quantity of deceptive and false information is steadily increasing (Song and Zahedi 2007). Online medical information is fraught with credibility issues, in which scam initiators rely on social expectations and psychological persuasion techniques to target patients (Garrett et al. 2019). For example, with prescriptions and doctor recommended medications, fraudulent actors often pose as online sources of information and products, resulting in counterfeit drugs and misinformation (Song and Zahedi 2007). Yet, detecting non-credible health-related content is not easy, even for humans (Li et al. 2019).

This presents a problem for users and social listening platforms alike as they try to assess health information online. Non-credible content provides a false signal for social listening tasks. Recognizing and eliminating content with low credibility during the social listening process will provide higher quality information in a more efficient manner. Fortunately, our extension to Activity Theory provides insight into signals that can help identify non-credible information. First, because authoritative and trustworthy authors (and similarly sites/channels) are less likely to link to non-credible content (Gyöngyi et al. 2004), links between authors may be used to infer credibility, similar to inferring relevance as illustrated in Figure 4. However, these methods require seed information about credibility, which is not always readily available (Abbasi et al. 2012). For this, we may look to group norms in the community.

In our Reddit case, a moderator explains the group norm for identifying non-credible authors: “If you are trying to vouch for someone simply leave a ‘+’ a positive experience, ‘-’ a negative [experience], and ’XXX’ stay away, bad friend.” This vernacular was observed frequently. For example, a user posted a new thread with pictures of opioids designed to entice purchasing, but subsequent responses from other uses simply stated, “xxx bad friend.” The phrase “bad friend” was observed numerous times as the vernacular chosen by the community to warn others the person was looking to harm or scam people. Without the context of understanding the “bad friend” warnings, the above post may have fit the profile of task-relevant information about an especially potent batch of opioids hitting the market, whereas in actuality it is likely a scam artist over-promising the potency of his drugs. An effective and efficient social listening platform must be able to utilize characteristics of communication activities propagating through online platforms and the multiplex relationship between them to evaluate information credibility.

### 4.5 Meta-Requirement 3: Navigate Complexity of the Online Channel Landscape

The final theme of contextual information quality that guides our meta-requirements is that of completeness. Data that is only a subset of the complete set of relevant data can lead to incorrect assumptions and interpretations (Lee et al. 2002; Todoran et al. 2015). A related concept is richness, in that the data that provides deep contextual detail is more useful (Zheng et al. 2013). Users and providers of online health content generate an enormous amount of information across a variety of channels, making it difficult to navigate and analyze (Chung et al. 2005). Existing social listening techniques often assume topical locality, which posits that similar content is likely to be well linked together (Davison 2000), making them prone to converging on local optima, getting trapped in pockets within the online landscape and missing large quantities of relevant data. Many healthcare information providers are in competition with other websites and forums, with sponsorship motivations to avoid crosslinks (Szalavitz 2011). Online communities tend to create pockets of communication that may be isolated from other, related, and relevant content (Bergmark et al. 2002). These health communities may span multiple communication channels, fragmenting and dispersing important content across the channel landscape, switching between relevant and irrelevant information within conversation threads, and creating many one-directional links, creating significant difficulties for navigation (Boudry 2015; Jami Pour and Jafari 2018).

In the opioid context, online resources come and go rapidly as sites get shut down and users congregate to discuss drugs in a new location, with sparse (or no) direct connections between these, as users want to avoid authorities as they move to new communication hubs (Ladegaard 2019). In order to reach these pockets of relevant content, a social listening platform may have to traverse links to intermediate sites that seem less relevant to the social listening task, but actually, eventually lead to more relevant information. To illustrate this phenomenon, we used detailed clickstream data from Comscore to analyze browsing activity of users that visited the r/opiates subreddit between 2012 and 2018. Comscore recruits and pays a representative panel of internet users to install an apparatus which records and reports their internet browsing behaviors. By analyzing the clickstream collected by Comscore, we identified all users who visited r/opiates at some point during their tenure on the panel, and compiled the top 100 sites visited by this group of users (as measured by percentage of time spent on the site). As expected, many of the sites were drug related and likely relevant to the opioid listening task.

However, as Figure 5 depicts, direct hyperlink connections between many of these sites are sparse, particularly among many of the confirmed drug-related sites (Park et al. 2018). Without consideration of the potential for certain irrelevant URLs to lead to relevant content, a social listening artifact would have a difficult time traversing between the relevant content. However, Bitcointalk.org emerges as an important node that forges connections between drug-related sites – in fact, it is the most central node in the entire graph of sites (nodes are sized by betweenness centrality, although bitcointalk.org was the most central for a variety of measures). This has face validity, as bitcoin is frequently used to pay for illegal drug transactions online (Foley et al. 2018). Without traversing links to/from Bitcointalk.org which might otherwise be considered irrelevant based on content alone, a social listening platform might have difficulty navigating the channel landscape to pockets of relevant content, instead converging on local optima and missing significant amounts of important information. In order to efficiently gather relevant public health information, a social listening platform must be able to navigate the complex channel landscape that is the reality of the online platform environment.

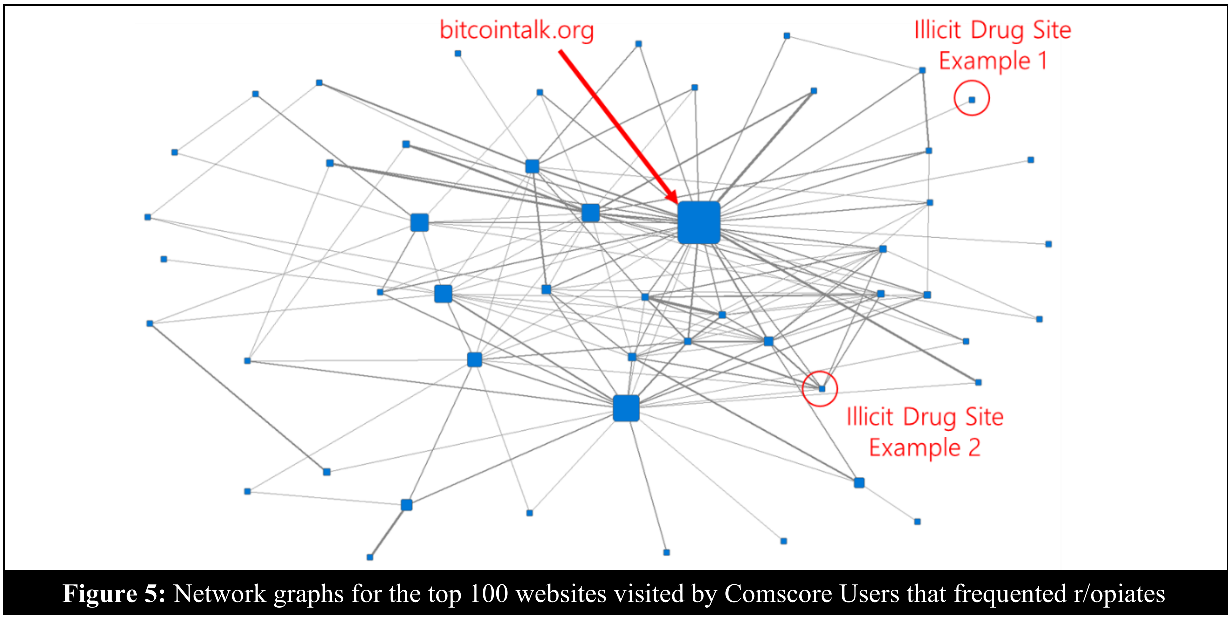


**Figure 5:** Network graphs for the top 100 websites visited by Comscore Users that frequented r/opiates

## 5 The Healthsense Artifact

We follow the requirements and design elements identified through Activity Theory and the contextual information quality literature to develop HealthSense, a social listening platform for public health. Our system is comprised of three modules, each of which address a separate meta-requirement: evaluating the 1) relevance and 2) credibility of content while intelligently 3) navigating the complex channel landscape of online platforms in order to gather timely, granular, and actionable information for use in public health informatics. All three modules leverage state-of-the-art graph neural networks in their construction. Recently, graph neural networks (GNNs) have garnered considerable attention for their ability to incorporate graph message parsing into robust machine learning architectures (Wu et al. 2021). GNNs' ability to parsimoniously consider nodes, edges, features, and hierarchical graph structures in the convolution process (Wu et al. 2021) aligns with the extension of Activity Theory to consider the propagation of information across multiplex interactivity relationships among communication activities in online platforms. From a computational design perspective (Padmanabhan et al. 2022; Rai 2017), our HealthSense instantiation makes the following methodological contributions:

- *Couple graph propagation with state-of-the-art GNNs*. This allows us to incorporate credibility information from across millions of nodes (via graph propagation), with limited seed labels, in unison with GNNs that are better suited for message-passing across thousands of nodes (i.e., localized focal node-level graphs) in relatively less sparse-label environments (Bojchevski et al. 2020; Dai et al. 2022). There has been limited work at this intersection (Bojchevski et al. 2020). Moreover, we employ multi-level bi-directional graph propagation to better capture the multiplex of activities across the network.

- *Propose bi-relational edge-enriched node embeddings*. We use domain-adapted feature-based classifiers to derive important node and in/out-bound edge information used by the GNNs for assessing relevance and tunneling potential, and represent these parsimoniously in an embedding that considers different relation types, and node/edge characteristics.

We discuss these novel aspects of HealthSense in the remainder of the section, and subsequently use benchmarking and ablation analysis to evaluate them, and HealthSense as a whole.

The Relevance Assessment Module (RAM) evaluates both topic and sentiment of online content in order to determine its relevance to the social listening task. The primary signals used are linguistic features of the information content surrounding a link to target content. These are evaluated within the context of latent author tendencies discovered through analysis of other authored content, and relationships to other authors and content. Relevance cues also include context-specific lexicons, libraries, and dictionaries which capture group norms from the communities within which the content was created. Multiple classifiers are created to incorporate varying structures and behaviors across online channels and used as input for the GNN.

The Credibility Assessment Module (CAM) evaluates the credibility of online content to avoid collection of non-credible information in the social listening task. CAM uses graph propagation-based features (in a similar manner as TrustRank (Gyöngyi et al. 2004)) as input for a GNN-based credibility classifier. Document, author, and site-level graphs are utilized to evaluate potential content. Trust for each graph is seeded using various signals according to group norms of how credibility is determined within various communities.

The Landscape Assessment Module (LAM) evaluates links among online platforms in order to identify content which may not contain relevant information itself, but may lead to further relevant content as document nodes in the graph are traversed. This counterbalances the hyper-focus of RAM and CAM on collection of relevant, credible content, and is critical as pertinent information often occurs in sparsely connected pockets. LAM utilizes a GNN to learn patterns from subgraphs from a training set of nodes known to lead to relevant or irrelevant information.

The HealthSense system is comprised of the RAM, CAM, and LAM modules, as summarized in Figure 6. The system begins with seed content URLs and propagates through links to other online content. Each candidate URL is first evaluated by the CAM module and rated with regard to its estimated credibility. Each URL which exceeds a specified threshold credibility is passed to RAM which ranks all current candidate URLs based on expected relevance to a given social listening task based on topic and sentiment information. Finally, the LAM module evaluates all URLs below a specified relevance threshold and scores them based on their likelihood of leading to further relevant content. The system then retrieves candidate URLs from a queue in order of combined RAM and LAM scores. As new URLs are collected, scores for all remaining candidate URLs are updated to incorporate new information. In this way, HealthSense gathers the most useful information for a given social listening task in a highly efficient manner.

As depicted in the center of Figure 6, each module's GNN employs a relational graph convolutional network (Kipf and Welling 2017; Schlichtkrull et al. 2017). Let $G$ represent the graph of collected and in-queue document nodes $V$, with $E$ signifying the set of edges (in and outlinks). For each node $v$ with neighbors $N(v)$, each layer $k$ of the graph convolutional network feeds forward a node embedding $\mathbf{h}_v^k$ (i.e., a feature vector) by averaging the neighbor nodes' information and passing it through a neural network. This process is repeated for each $v$, across the $K$ layers of the graph convolutional network, with each subsequent layer pulling in neighbor information from one further hop (i.e., neighbors' neighbors, and so on). Although the proposed node embeddings can provide invaluable micro-level insights, we aggregate node embeddings into a graph embedding such that CAM, RAM, and LAM predict credibility, relevance, and tunneling potential for a given document node as a graph classification problem to allow better consideration for macro-level in/out-link information. To ensure that node and graph embeddings are learned as part of an end-to-end learning strategy, the graph convolutional network's loss function uses final binary class labels from a small training set (relevance for RAM, credibility for CAM, and tunneling potential for LAM). The initial input node embeddings for each $v$ are:

$$\mathbf{h}_v^0 = \mathbf{x}_v \qquad (1),$$

where $\mathbf{x}_v$ is the feature vector of node $v$. For each subsequent layer, where $k \geq 1$:

$$\mathbf{h}_v^k = \sigma\left(\sum_{\mathrm{r}\in R}\sum_{u\in N(v)} \mathbf{W}_k^r \frac{\mathbf{h}_u^{k-1}}{|N_r(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1}\right) \qquad (2),$$

where $\sigma$ is an activation function for non-linearity in the node embeddings. $\mathbf{W}_k^r$ and $\mathbf{B}_k$ are the trainable model weight parameters for the node embedding dense layers, used to balance how much new neighborhood information should be aggregated across each $u$ neighbor of node $v$ in the latest convolutional layer $k$, versus existing node embedding information (from the prior layer). $R$ denotes the set of relation types in the graph – in our case, there are two distinct types: inlinks and outlinks. Hence, $N_r(v)$ are the inbound or outbound neighbor nodes. As we employ graph convolutional networks designed for undirected graphs (Wu et al. 2021; Kipf and Welling 2017), adding a relational mechanism enables us to account for differences in propagation information across in versus out-bound links (Schlichtkrull et al. 2017). That is, the graph convolutional network includes separate weights for each relation type at each layer.

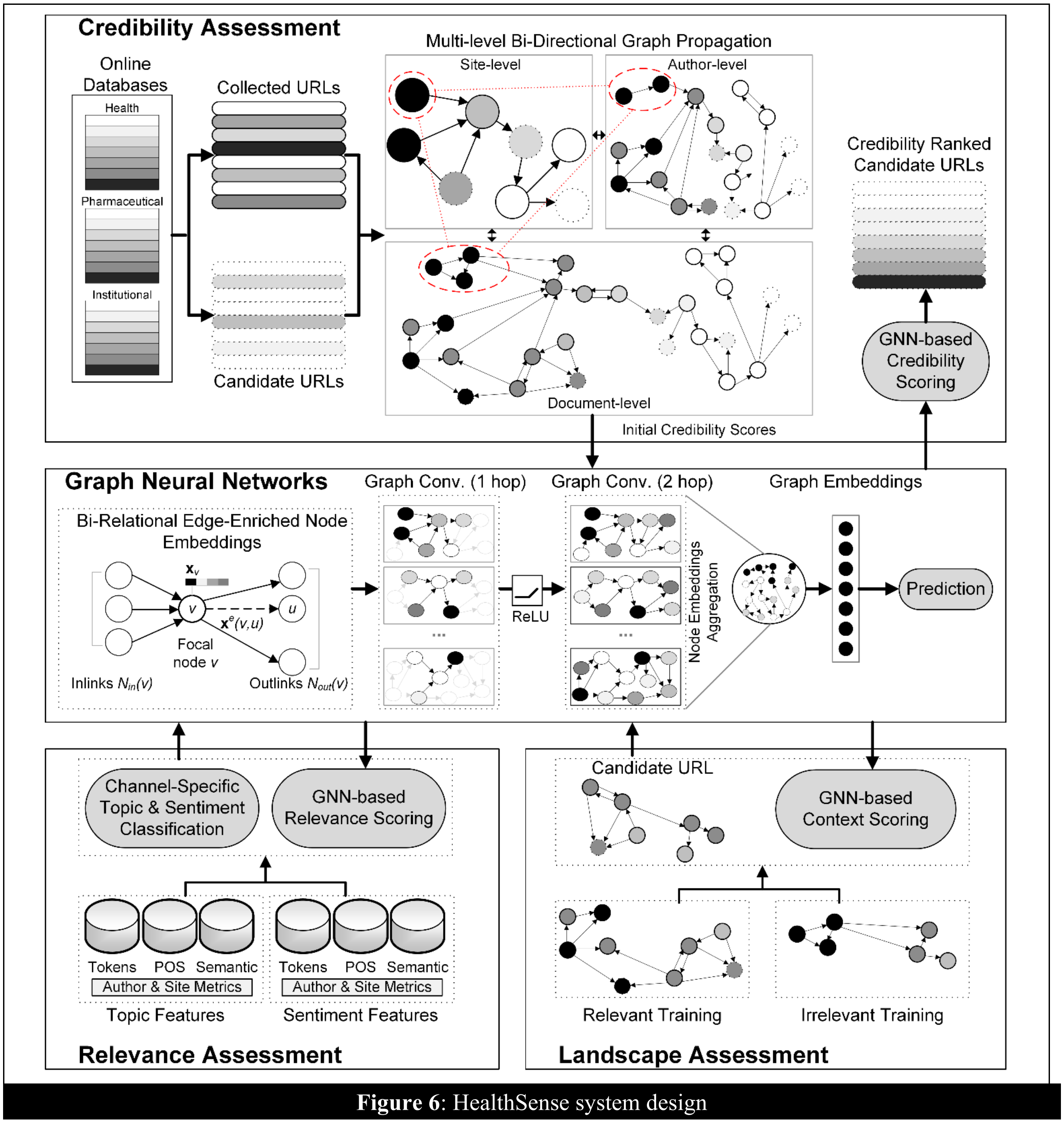


**Figure 6**: HealthSense system design

One limitation of the node embedding formulation in equation (2) is that it doesn't consider edge feature vectors. For relevance and landscape assessment, where we want to decide whether to retrieve a given URL, the anchor text around a link might contain important edge-specific information between nodes (Fu et al 2012). To incorporate in/outbound edge feature vectors $\mathbf{x}_{(v,u)}^e$ between nodes $v$ and $u$, we can tweak our formulation to an edge-conditioned node:

$$\mathbf{h}_v^k = \sigma\left(\sum_{\mathrm{r}\in R}\sum_{u\in N(v)} \mathbf{W}_k^r \frac{\mathbf{h}_u^{k-1}\,\mathbf{F}^r\mathbf{x}_{(v,u)}^e}{\sqrt{|N_r(v)||N_r(u)|}} + \mathbf{B}_k\mathbf{h}_v^{k-1}\right) \qquad (3),$$

where $\mathbf{F}^r$ is the weights from a single dense layer neural network learned offline across all edge vectors for relation type $r$ in $E$. To normalize the added dense edge feature representation signified by $\mathbf{F}^r\mathbf{x}_{(v,u)}^e$, we take the dot product between the neighbor edge and node vectors.

As noted, the three respective GNNs treat node credibility, relevance, and tunneling potential prediction for each node $v$ as a graph classification problem. Whereas node embeddings could be aggregated by averaging or summation to construct a simple graph embedding, such an approach would lose important node heterogeneity information. Instead, we employ non-linear aggregation using multi-layer perceptrons to calibrate node importance. Formally, given $\mathbf{h}_v^k, \mathbf{h}_1^k, \mathbf{h}_2^k \ldots \mathbf{h}_n^k \in H^k$ denotes the set of node embeddings associated with node $v$ and its neighbors at the last layer $k = K$, the graph embedding $g_G$ uses a two-layer multi-layer perceptron where the first layer weights $\mathbf{W}_{l_1}$ are a non-linear accumulation of the node embeddings and the second layer $\mathbf{W}_{l_2}$ maps this information to a dense representation using a non-linear (ReLU) activation function:

$$g_G = \sigma\left(\mathbf{W}_{l_2}\sigma\left(\sum_{\mathbf{h}_i^k\in H^k} \mathbf{W}_{l_1}\mathbf{h}_i^k\right)\right) \qquad (4),$$

For CAM and LAM, $H^k$ includes $N_r(v) \cup N_r(u)$ to allow inclusion of two-hops of neighbors, but for RAM $H^k$ encompasses $N_r(v)$ since the target node is more crucial for relevance assessment. All three GNNs use the ReLU activation function for node embedding layers and binary cross-entropy for loss. Finally, a single dense layer is used to predict the downstream task (i.e., relevance, credibility, and/or tunneling potential):

$$\hat{y} = \mathbf{W}_f * g_G \qquad (5),$$

The node and edge feature vectors $\mathbf{x}_v$ and $\mathbf{x}_{(v,u)}^e$ for the target node, while important throughout for the GNNs to infer credibility, relevance, and tunneling potential, become even more critical early on due to the sparse nature of the graphs (i.e., there is highly incomplete information about neighbors in the earliest listening phases, as few nodes have yet been collected). Consequently, for each node $v$, the RAM, CAM, and LAM modules leverage graph propagation or feature-based classification methods to compute initial credibility, relevance, and context scores/features which are used as the node and edge vectors for the GNNs used to derive the final $C(v)$, $S(v)$, and $L(v)$ classification scores. In the remainder of the section, we describe how these crucial initial features/scores are derived.

### 5.1 Relevance Assessment Module (RAM)

RAM evaluates each candidate URL $v$ to be retrieved to determine its potential relevance to the social listening task. For each $v$, the initial relevance score used as input feature vectors for the GNN are derived based on text classifiers trained to determine the topic and sentiment of content surrounding the URL in parent documents (those that contain links to the candidate URL). This module represents the most direct evaluation of whether the content for collection is relevant to the task at hand. For example, if the social listening objective were to be related to detecting adverse drug events, the classification models in RAM would be trained to target content expressing negative sentiments about experience with a prescription medication. To account for channel-specific communication patterns (Smith et al. 2012; Waterloo et al. 2018), topic and sentiment classifiers are trained on separate labeled corpora categorized into four channels: web, social, forum, and blog, resulting in 8 total binary classifiers.

As depicted in Table 3, each classifier employs a wide variety of features, including fixed word unigrams, bigrams, and trigrams, as well as various linguistic features and lexicons designed to capture group norms. Part-of-speech (POS) and POS-word n-grams are derived using the Stanford tagger and the CMU ARK tool in order to account for channel-specific language usage (Manning et al. 2014). N-grams are also coded with semantic features from a variety of lexicons to capture more generalizable language patterns (Baccianella et al. 2010). These include general and channel-specific entity tags including emojis, abbreviations, and slang terms (Zimbra et al. 2018). Entity tags are curated from domain-specific lexicons and lexical thesauri. For instance, if the domain of interest is post-marketing surveillance of pharmaceutical drugs, example entity tags might include <*drug*>, <*condition*>, <*symptom*>, and <*treatment*>. The POS and semantic entity tags both help interpret the meaning of messages by abstracting and categorizing community-specific language characteristics into a common scheme allowing analysis. Finally, semantic sentiment tags are derived from existing sentiment lexicons containing polarity scores for over 100,000 terms (Baccianella et al. 2010). Terms are labeled with one of seven possible sentiment labels from <*strong neg*> to <*strong pos*>, based on average sentiment scores across its various senses. Several author metrics are employed to account for the impact of author roles and tendencies on potential topic or sentiment relevance of a document's outlinks. These pertain to the topic and sentiment relevance of other known authored documents, and number/percentage of author in/out links as well as topic or sentiment relevant ones.

**Table 3**: Examples of N-Gram Features Used In Topic and Sentiment Relevance Classifiers

| Feature Category | Example Text Representation | Key Topic Feature Examples | Key Sentiment Feature Examples |
|---|---|---|---|
| Word | I used to hate smoking until I started taking Chantix ☹. | smoking, started taking, Chantix | hate, smoking, Chantix |
| POS | PRP VBD TO VB NN IN PRP VBD VBG NNP : : LRB | NN, NNP, VBG NNP | VB, VB NN |
| POS-Word | PRP\|I VBD\|used TO\|to VB\|hate NN\|smoking IN\|until PRP\|I VBD\|started VBG\|taking NNP\|Chantix :\|: :\|- LRB\|( | NN\|smoking, NNP\|Chantix | VB\|hate, :\|: :\|- LRB\|( |
| Semantic Entity Tags | I used to hate <habit> until I started taking <drug> <frown face>. | <habit>, taking <drug>, <frown face> | |
| Semantic Sentiment Tags | <neu> <neu> to <strong neg> <neu> until <neu> started <neu> Chantix <strong neg>. | | <strong neg>, <strong neg> <neu> |
| Author and Site Content | Number of known prior authored documents, number and percentage of topic/sentiment relevant documents | | |
| Author and Site Linkage | Number of total links, number and percentage of in/out links, number and percentage of topic/sentiment relevant in/out links | | |

To refine the extracted feature space for faster relevance assessment in time-sensitive environments and improve classification performance, the attribute space is ranked and filtered using a feature subsumption approach based on the information gain heuristic (Riloff et al. 2006). For each feature *f*, information gain IG(*f*) is calculated based on entropy reduction provided by that feature in isolation. Only features with IG(*f*) greater than a defined threshold were retained for use in each classifier. Further, no higher-order n-gram g was retained unless it provided higher information gain than all q of the lower-order n-grams it contained (i.e., $\mathrm{IG}(g) > \mathrm{IG}(g_i)\ \forall i \in [1,q]$).

Each of the 8 channel-specific topic/sentiment models was trained using a binary linear SVM classifier (SVMperf) (Joachims 2006). These classifiers were trained on a set of manually labeled content known to be relevant or irrelevant from a topic or sentiment perspective. To ensure that sentiment was only evaluated with regard to the target topic, the sentiment classifier only used features extracted from defined windows surrounding relevant topic keywords. Topic and sentiment relevance scores from the anchor text surrounding in-bound links from each $u$ towards target document $v$ are used to construct $\mathbf{x}^{e}_{(v,u)}$ edge feature vectors in the GNN. The $\mathbf{x}_v$ node feature vectors are one-hot encoded with the index value corresponding to that node set to 1, and all other values in the vector 0. For each $v$, the GNN computes a relevance score $S(v)$.

### 5.2 Credibility Assessment Module (CAM)

The credibility assessment module (CAM) is intended to reduce the intake of low-credibility content, including medical spam and phishing documents, which can pose major information quality concerns if unmitigated. In order to accomplish this objective, CAM uses weighted multi-layer bi-directional graph-propagation to construct a credibility value for each node. These values are input as the $\mathbf{x}_v$ node feature vectors in equation (1) of the credibility GNN. Our graph propagation-based credibility features method addresses several limitations associated with existing propagation techniques. Although prior studies have relied on only document-level graphs (Diligenti et al. 2000; Farag et al. 2018), our extension to Activity Theory proposes that relationships between content, channel, author-level information provides important context. Therefore, the graph propagation in CAM incorporates inter-related graphs from each level. These graphs are seeded with credibility information from many online databases that maintain site, document, and/or author-level assessments. We incorporated many databases which are focused on guiding consumers to and accrediting credible online health information, such as the Medical Library Association (mlanet.org), the Health on the Net Foundation (hon.ch), the National Association of Boards of Pharmacies (nabp.pharmacy), and LegitScript (legitscript.com).

Authorship credibility is an emerging area of focus with the proliferation of cyber deviance and fake news on social media platforms (Viviani and Pasi 2017). For web and blog channels, if authorship information is present, initial author credibility is inferred from site and page-level credibility scores. For forums, such information is derived from community-level measures such as up/down votes on postings. For social media platforms, initial credibility is derived from existing databases or computed, using metrics such as TwitterRank (Weng et al. 2010).

While information quality from the databases we use to seed credibility is high, coverage is limited – typically less than 5% of all domains, URLs, and users. For instance, most of the aforementioned domain-level databases contain information for only thousands out of the millions of medical websites, pages, and authors existing online (Abbasi et al. 2012). To extrapolate this information to each URL evaluated by HealthSense for collection, we use graph propagation methods to project credibility information to unknown nodes (Gyöngyi et al. 2004). CAM utilizes a multi-level bi-directional graph-based algorithm that propagates over site, document, and author-level hyperlink graphs to compute the credibility of all nodes in the multi-level graph, including the candidate URLs. Bi-directional propagation allows for more efficient and effective usage of existing credibility information in sparse-graph situations where only retrieved and candidate nodes are available. The algorithm employs trainable parameters for initial versus propagated credibility, inbound versus outbound propagation, and cross-layer propagation.

In the document-level graph, a "document" refers to content with a unique URL, as in the case of a web page or blog page. For forum posts or social media messages with unique pages (URLs), a document would be the page associated with that particular post or message. However, for some forums or social media, it might be the thread page containing multiple posts/messages. Each node $v$ is assigned an initial credibility score $C'_I(v)$ depending on database coverage and node type (e.g., site, document, author). Regarding sites and pages/documents, depending on the database used to seed credibility, information may be available at the individual page level, or only at the site level. If only site-level information is available, all pages in the document graph are assigned the same credibility label as the site. If the converse is true and only page-level credibility is known, the site is assigned a credibility score equal to the average of all known pages belonging to the site domain. Once the initial $C'_I(v)$ has been computed for each of the collected nodes and candidate URLs, credibility scores are propagated to all other nodes in the graph.

The credibility score for any node $v$ in the multi-level graph is computed as:

$$C'(v) = \alpha C'_I(v) + (1-\alpha)\left(\sum_{u_i \in N_i(v)} \beta \frac{C(u_i)}{|N_o(u_i)|} + \sum_{u_o \in N_o(v)} (1-\beta) \frac{C(u_o)}{|N_i(u_o)|}\right) + \gamma\left(\sum_{m \in H(v)} \frac{C(m)}{|H(v)|}\right), \quad (6)$$

where $i$ and $o$ denote the inlink and outlink edge relation types, and $N_i(v)$ and $N_o(v)$ are the sets of inlinks and outlinks of $v$, respectively. The $\alpha$ parameter controls the relative weights of initial versus computed credibility scores, and weights of inlinks versus outlinks are determined by the $\beta$ parameter. For any candidate URL, the set of outlinks is empty (i.e., $|N_o(v)| = 0$), since the URL has not yet been retrieved and its outbound links are as yet unknown. For the site and page-level graphs, in/out links denote hyperlinks pointing from one node to another. For the author-level graph, these links represent page-level hyperlinks between pages co-authored by respective authors. That is, if page $a$ points to page $b$, in the author-level graph, each author node from $a$ would have a link to author nodes in $b$. Within the multi-layer graph, cross-layer links (i.e., those between site and page nodes, pages and authors, and sites and authors), do not have directionality since these are not hyperlinks pointing from one site (page) to another, or directed between-author links. Hence, $H(v)$ are the set of links of $v$ with nodes from the other two layers in the graph. The parameter $\gamma$ determines the relative weight given to cross-layer links. The GNN uses the output of the graph propagation process described, $C'(v)$, as an input in $\mathbf{x}_v$. It is important to reiterate that the credibility scoring GNN does not use edge-conditioned node embeddings, but rather the embedding described in equation (2), as no topic and sentiment information is yet available for edge vectors (that is done later in RAM). CAM is intended to efficiently make an initial determination of document nodes without in-depth content analysis. As noted earlier, and later demonstrated empirically, our proposed multi-level bi-directional graph propagation approach used in concert with a GNN for credibility assessment, allows markedly better listening capabilities in the form of higher precision and recall, allowing efficient prioritization of relevant information to allow for timely collection.

## 5.3 Landscape Assessment Module (LAM)

The landscape assessment module (LAM) in HealthSense determines the potential for any given candidate URL to lead to additional relevant information by analyzing a labeled graph of all collected and candidate URL nodes. The context scoring GNN uses a training set encompassing subgraphs for documents with class labels indicating whether they led to relevant pages within $y$ hops. For a candidate URL $v$, it generates a context score $L(v)$. As input for the GNN's edge-conditioned node embeddings (see equation (3)), each $\boldsymbol{x}_v$ comprises features representing the node's channel source, estimated topic relevance, and estimated sentiment relevance. Channel sources are represented in $\boldsymbol{x}_v$ as a two-bit encoding. Average estimated topic and sentiment relevance scores across all inlinks of $v$ are provided by the RAM module. Additionally, the topic and sentiment relevance scores from anchor text surrounding in-bound links from each $u$ towards target document $v$ are used to construct the $\mathbf{x}^{e}_{(v,u)}$ edge feature vectors in the GNN.

### 5.4 Prioritized Collection Based on RAM, CAM, and LAM

Bringing all modules together, candidate URLs are collected based on their rank defined as

$$P(v) = \begin{cases} 0 \; if \; C(v) < T_C \\ S(v) \; if \; C(v) \geq \; T_C \; and \; S(v) \geq \; T_S \\ S(v) + L(v) \; if \; C(v) \geq \; T_C \; and \; S(v) < \; T_S, \end{cases} \tag{7}$$

where $S(v)$, $C(v)$, and $L(v)$ represent the aforementioned GNN-classifier scores from the RAM, CAM, and LAM modules, respectively, and $T_C$ and $T_S$ represent minimal credibility and relevance thresholds. Note that URLs below a certain credibility score are given the lowest possible priorities, although credibility scores are updated periodically based on more complete graph information and may rise above this threshold in any given update. This allows the RAM module to only be run on a subset of URLs. Likewise, the LAM module is only run for URLs that fall below a certain relevance threshold. URLs with $L(v) \geq 0$ are those with promising landscape contexts, and this score can help boost the relevance of a borderline candidate URL. The hierarchical flow of module execution for each URL significantly improves runtime for the HealthSense system. For simplicity, we use 0.5 as the threshold for $T_C$ and $T_S$, as this is also the standard classification cut-off used by our GNNs during the training phase. Tuning of these thresholds could lead to further improvement.

## 6 Applications and Evaluation

In order to evaluate HealthSense, we constructed a large-scale test bed and used it to explore two public health 3.0 tasks: (1) post-marketing drug surveillance (PMDS) and (2) synthetic opioid batch surveillance. As discussed in our motivating case above, the opioid listening task requires the detection of positively valenced content related to opioid usage in near real-time in order to be helpful to emergency responders, ER physicians, and law enforcement. For PMDS, although pharmaceutical companies are required to perform rigorous testing to ensure the safety of drugs before they may be sold to consumers, inevitably issues of adverse reactions or other unintended consequences occasionally arise after a drug is on the market. Because of the high social and monetary costs associated with these incidents, pharmaceutical companies, as well as other stakeholders such as regulators, watchdog agencies, and even healthcare hedge funds, are very interested in detecting such issues as early as possible (Brewer and Colditz 1999; van Grootheest et al. 2003). Effective, efficient PMDS relies upon the timely identification of signals that may suggest adverse reactions or other issues. Content posted on online platforms as consumers and providers discuss the drugs in question, particularly with a negative sentiment, provides a rich source to monitor for such signals. Accordingly, we use HealthSense to identify negatively valenced content from online platforms for use in PMDS tasks.

For both tasks, PMDS and opioid listening, we evaluate the amount of relevant content collected at various milestones by HealthSense compared to leading benchmarks. For the PMDS task, we also perform a field study, including a user experiment and disproportionality analysis, in order to evaluate the practical value of HealthSense. In the following sections, we describe the test bed data, comparison methods used, and detailed results of our field study and evaluation.

### 6.1 Testbed Construction

HealthSense is a social listening tool which intelligently senses and prioritizes collection of content most useful for a specified task. This intelligent sensing provides the efficiency imperative for public health tasks – the same information could be obtained by a simple crawler collecting all links, although in a much longer period of time and buried in a mountain of additional useless content. To evaluate the effectiveness and efficiency of HealthSense in gathering task-relevant information, we developed a test bed of all information that could be collected by a simple crawler given unlimited time. From a set of 100 seed URLs of health and drug-related sites, a simple crawler was used to collect the content of over 37 million distinct URLs from websites, forums, blogs, and social networking sites. Following prior work (Pant and Srinivasan 2005; Srinivasan et al. 2005; Fu et al. 2012), to determine the relevancy of each URL to the PMDS and opioid surveillance tasks, a gold-standard classifier too computationally expensive to be used for data collection was trained based on a training set of 2000 relevant and 2000 irrelevant pages manually labeled by two domain experts for each task.[1]

[1] In order to validate the results of the gold-standard classifier for evaluation, we conducted additional tests (available upon request) noting the performance to be statistically indistinguishable from that measured on the manually labeled URLs.

Consistent with prior online health studies, the annotation protocol involved two steps (Cameron et al. 2013; Leaman et al. 2010). First, an appropriate ontology was used to gauge the presence of relevant entities in the online document (i.e., page). This part involved assessing whether a page indeed referenced an entity of interest for that particular testbed. Next, where a relevant entity existed in the page, the valence of the entity mention was assessed. For the first step, in regards to the opioid test bed, annotators used the drug abuse ontology framework (Cameron et al. 2013; Nasralah et al. 2020) to identify entities closely related to opioids. In the case of PMDS, the entity lexicon was comprised of terms from UMLS, SIDER, idiomatic expressions, and the consumer health vocabulary (Leaman et al. 2010). For the second step, annotators of both testbeds used guidelines from the online drug epidemiology literature (Cameron et al. 2013), coupled with the general guidelines for valence annotation (Wiebe et al. 2005) used in many prior IS/online studies. The former are well-suited for health domain-specific valence annotation whereas the latter offer guidelines on assessment of valence for an array of subtle private states manifesting in user-generated text, such as "opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments," (Wiebe et al. 2005, p. 168). As noted, for the opioid (PMDS) listening task, positive (negative) valence pages were deemed relevant.

**Table 4:** Test Bed Overview

| Source | Total | PMDS Task | | | Opioid Task | | |
|---|---|---|---|---|---|---|---|
| | | Relevant | Irrelevant | % Relevant | Relevant | Irrelevant | % Relevant |
| Website | 14,281,509 | 1,582,096 | 12,699,413 | 12.46% | 860,012 | 13,421,497 | 6.41% |
| Blog | 4,111,213 | 387,827 | 3,723,386 | 10.42% | 624,424 | 3,486,790 | 17.91% |
| Forum | 3,553,363 | 488,839 | 3,064,524 | 15.95% | 367,418 | 3,185,945 | 11.53% |
| Social Network | 15,118,657 | 2,410,606 | 12,708,051 | 18.97% | 1,213,490 | 13,905,167 | 8.73% |
| **Total** | **37,064,742** | **4,869,368** | **32,195,374** | **13.14%** | **3,065,343** | **33,999,398** | **9.02%** |

On average, these pages comprised 1,687 words/tokens. The annotators labeled an additional 12,000 test pages as relevant/irrelevant to PMDS and opioids. Following best practices (Abbasi et al. 2018), the experts underwent two rounds of discussion to resolve differences after independently annotating 100-page samples. They each then independently labeled all training and test cases, meeting after every 1000 instances to resolve differences. Before resolving differences, Cohen's Kappa across all 16,000 instances was 0.95 and 0.93 for PMDS and opioid tasks, respectively, representing very good inter-rater agreement. Consistent with prior work (Fu et al. 2012; Menczer et al. 2004), this SVM classifier using over 100,000 n-gram features (n = 1, 2, 3) was designed to be highly accurate but is unsuitable for time-sensitive tasks. Evaluation on an independent set of 12,000 test pages found it to have relevance classification accuracies of 97.2% and 95.3% on the PMDS and opioid tasks, respectively. The classifier was applied to the entire set of 37 million URLs to determine the relevancy of each, as summarized in Table 4.

**Table 5**: Baseline and benchmark comparison methods

| Method | Reference | Description |
|---|---|---|
| ***Graph Neural Networks*** | | |
| Heterogeneous graph attention network (HGAN) | (Wang et al. 2019) | Each candidate URL's relevance score is calculated using a GNN relevance classifier that takes into account node and link type information and hierarchical attention. Node types are defined by the 4 channels and topic/sentiment scores derived using BERT (i.e., 16 types). Link types are in- and out-links. |
| Het. graph neural network (HGNN) | (Zhang et al. 2019) | Each candidate URL's relevance score is calculated using a GNN relevance classifier that takes into account node type information (similarly to HGAN). |
| Graph attention network (GAT) | (Veličković et al. 2018) | Each candidate URL's relevance score is calculated using a GNN relevance classifier that implicitly captures the weights for all edges with parametric weights. |
| GraphSAGE (GSage) | (Hamilton et al. 2017) | Each candidate URL's relevance score is calculated using a GNN relevance classifier that samples neighboring nodes and assumes they contribute equally. |
| Graph conv. network (GCN) | (Kipf and Welling 2017) | Each candidate URL's relevance score is calculated using a GNN relevance classifier that explicitly captures the weights for all edges with fixed non-parametric weights. |
| ***Focused crawlers utilizing hyperlink graph information*** | | |
| Graph-Based Sentiment (GBS) | (Fu et al. 2012) | Candidate URLs are ranked based on their topic and sentiment composition across the hyperlink graph, relative to known relevant and irrelevant URLs |
| Hopfield Net (HFN) | (Chau and Chen 2007) | Each candidate URL has weighted links from inbound nodes in a single-layer neural network, where weights, activation, and loss are handled using feed-forward, back propagation based on actual relevance once collected |
| Context Graph Model (CGM) | (Diligenti et al. 2000) | Candidate URLs are ranked based on their classification scores using a series of Naïve Bayes classifiers each trained on documents that are exactly *n* hops away from relevant content, with *n*=0 indicating directly relevant content |
| ***Node embeddings learned using neural networks*** | | |

| | | |
|---|---|---|
| DeepWalk (DW) | (Perozzi et al. 2014) | Node random walk sequences over the hyperlink graph are input into a skip-gram (neural network with one hidden layer) to construct node embeddings used to rank candidate URLs based on cosine similarities with known relevant and irrelevant URLs |
| Node2vec (N2V) | (Grover and Leskovec 2016) | Node breadth-first and depth-first random walk samples over the hyperlink graph are input into a skip-gram model (neural network with one hidden layer) to construct node embeddings used to rank candidate URLs based on cosine similarities with known relevant and irrelevant URLs (i.e., similar to DW, but with breadth/depth control) |
| ***Focused crawlers utilizing link context*** | | |
| Keyword (KW) | (Aggarwal et al. 2001) | TF-IDF vectors for a select pre-defined list of keywords are extracted from the content of the training URLs and used to rank candidate URLs for collection based on vector cosine similarities with known relevant and irrelevant URLs |
| Vector-Space Model (VSM) | (Aggarwal et al. 2001) | TF-IDF vectors for all present tokens are extracted from the content of the training URLs and used to rank candidate URLs for collection based on vector cosine similarities with known relevant and irrelevant URLs |
| Naïve Bayes (NB) | (Pant and Srinivasan 2005) | N-grams are extracted from the content of the training URLs as features in a naïve Bayes classifier for ranking candidate URL relevance probability |
| BERT | (Yang et al. 2022) | Candidate URLs ranked based on relevance scores computed using a BERT-base model further fine-tuned on our relevant versus irrelevant training cases. |
| ***Baseline crawlers*** | | |
| Breadth-First Search (BFS) | (Chau and Chen 2003) | Each candidate URL is collected in a breadth-first queue from the list of seeds |
| PageRank (PR) | (Brin and Page 1998) | Each candidate URL is collected in decreasing order of PageRank computed from existing URL graph prior to collection |

### 6.2 Data Experiment

HealthSense was evaluated against a set of baselines and leading benchmark methods for social listening on both PMDS and opioid listening tasks. Because of potential sensitivity to seeding, each method was evaluated based on average performance in collecting URLs across 10 separate runs, each run seeded with a random subset of 200 seed URLs from a pool of 500 seed URLs (distinct from those used in test bed construction). Each benchmark method was trained as described in Table 5 to identify pertinent content based on the 4,000 expert-labeled URLs.

In Table 6 and Figures 7 and 8, we evaluated the performance of the HealthSense system and benchmarks by comparing precision, recall, and f-measure metrics. In Table 6, the AUC values in the first three columns are the areas under the curve for the f-measure, precision, and recall curves depicted in Figure 7 and 8 – a measure of the overall shape of the curves, with percentages closer to 100% indicating better performance. Note that due to space limitations, we only plotted the top two comparison methods from each category. The HealthSense system attained 75.1% of all relevant pages for the PMDS task within the first 5 million URLs collected, with a precision of 72.6%. These figures are more than double those of the best comparison method, HGAN. For the opioid surveillance task, HealthSense was 2.4x better than HGAN at 5 million URLs.

At 10 million URLs collected, HealthSense identified 99.9% of all relevant pages for both tasks, while HGAN lagged behind at between 53-67%. HealthSense outperformed other advanced GNN and graph analytics methods by even wider margins. As may be seen Figure 7, this advantage is not limited to any particular channel, with HealthSense performing consistently across website, blog, forum, and social networking formats. With regard to timeliness, it is important to note that server response represent the sole bottleneck for collection – algorithm runtimes are significantly shorter than the time spent waiting on requested pages. Therefore, the time spent in collection is a linear function of the performance noted in Table 6 and Figures 7 and 8. For instance, HealthSense can reach 75% recall after collecting 5 million pages – approximately 2.5x faster than HGAN, which would need to collect 12 million pages, and over 4x faster than a baseline BFS (22 million pages). Exact collection times are highly dependent on network and collection infrastructure.

It is clear from the results that HealthSense is extremely effective and efficient at prioritizing and collecting relevant information for social listening tasks, significantly better than existing state of the art methods. As we later demonstrate with our field study, these performance lifts translate into significant improvements in downstream listening tasks – e.g. allowing key stakeholders to identify important adverse events faster and more accurately. In order to demonstrate how theory-guided design elements in HealthSense impact its overall performance, we performed an ablation analysis (Padmanabhan et al. 2022). Consistent with prior machine learning-oriented design evaluation (Yang et al. 2022), we conducted a leave-out analysis at the module and individual component level, using ablation settings as described in Table 7. For module-level, we examined performance deltas attributable to excluding CAM and/or LAM modules from HealthSense. Connecting to Table 2, this category of ablation analysis is analogous to excluding the entire columns related to requirement (2) and/or (3). Component-level ablation focused on the impact of excluding specific elements related to individual cells labeled in Table 2.

The values depicted in Table 8 are the percentage degradation in performance when each component is excluded from HealthSense. Based on the first three rows (for each testbed), performance suffers significantly when excluding CAM or LAM, and even further if both are removed. Ablation of components (a) thru (j) highlight the importance of individual design elements (all paired t-test p-values < 0.01). The most impactful are elements i and j which relate to our extension to Activity Theory. This analysis demonstrates the significant value contained in multiplex interactivity relationships through which information propagates on online platforms.

**Table 6:** Percentage AUC Values and F-measure, Precision, and Recall at 5 and 10 Million

**PMDS Tasks**

| **Method** | | **AUC Values*** | | | **@5M** | | | **@10M** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **F-Meas** | **Prec** | **Rec** | **F-Meas** | **Prec** | **Rec** | **F-Meas** | **Prec** | **Rec** |
| HealthSense | | **44.7** | **37.9** | **91.8** | **73.8** | **72.6** | **75.1** | **65.8** | **49.1** | **99.9** |
| Graph Neural Networks | HGAN | *33.2* | *26.0* | *79.1* | *35.9* | *35.2* | *36.7* | *42.9* | *31.7* | *66.3* |
| | HGNN | 31.3 | 24.3 | 76.3 | 33.7 | 32.5 | 35.0 | 40.2 | 29.4 | 63.4 |
| | GSage | 29.8 | 22.8 | 74.7 | 32.0 | 30.0 | 34.2 | 37.9 | 27.3 | 62.0 |
| | GAT | 28.9 | 22.0 | 73.9 | 31.2 | 29.0 | 33.9 | 36.8 | 26.3 | 61.3 |
| | GCN | 28.1 | 21.3 | 73.1 | 30.5 | 27.9 | 33.5 | 35.7 | 25.3 | 60.7 |
| Hyperlink Graph | GBS | 32.2 | 25.1 | 75.9 | 33.9 | 33.5 | 34.3 | 38.4 | 28.6 | 58.7 |
| | CGM | 28.3 | 20.5 | 70.2 | 23.3 | 23.0 | 23.6 | 31.5 | 23.4 | 48.1 |
| | HFN | 21.9 | 15.1 | 58.0 | 14.9 | 14.7 | 15.1 | 22.2 | 16.5 | 33.9 |
| Node Embedding | N2V | 25.3 | 18.1 | 64.1 | 19.8 | 19.5 | 20.0 | 25.4 | 18.9 | 38.7 |
| | DW | 24.0 | 16.3 | 63.1 | 15.4 | 15.2 | 15.6 | 23.0 | 17.1 | 35.2 |
| Link Context | BERT | 25.4 | 17.5 | 65.5 | 16.8 | 16.7 | 16.9 | 25.0 | 18.7 | 37.7 |
| | NB | 24.5 | 16.8 | 64.2 | 16.2 | 16.1 | 16.4 | 24.1 | 17.9 | 36.8 |
| | KW | 21.2 | 14.5 | 56.8 | 13.2 | 13.0 | 13.3 | 19.9 | 14.8 | 30.4 |
| | VSM | 19.9 | 13.3 | 55.0 | 10.3 | 10.2 | 10.4 | 16.0 | 11.9 | 24.4 |
| Baseline | BFS | 25.5 | 19.6 | 62.5 | 24.8 | 24.5 | 25.1 | 26.5 | 19.7 | 40.4 |
| | PR | 19.9 | 13.9 | 52.9 | 12.8 | 12.6 | 12.9 | 18.8 | 14.0 | 28.6 |

**Opioids Task**

| **Method** | | **AUC Values*** | | | **@5M** | | | **@10M** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **F-Meas** | **Prec** | **Rec** | **F-Meas** | **Prec** | **Rec** | **F-Meas** | **Prec** | **Rec** |
| HealthSense | | **33.1** | **25.9** | **90.6** | **51.7** | **41.7** | **67.9** | **47.6** | **31.2** | **99.9** |
| Graph Neural Networks | HGAN | *21.3* | *14.5* | *73.8* | *21.9* | *17.5* | *29.2* | *24.8* | *16.1* | *53.8* |
| | HGNN | 19.8 | 13.2 | 72.5 | 20.3 | 15.7 | 28.6 | 22.4 | 14.3 | 51.8 |
| | GSage | 18.3 | 12.1 | 70.9 | 18.8 | 14.1 | 28.3 | 20.6 | 12.9 | 51.8 |
| | GAT | 18.1 | 11.9 | 70.4 | 18.5 | 13.8 | 27.9 | 20.3 | 12.7 | 51.2 |
| | GCN | 17.9 | 11.8 | 69.9 | 18.2 | 13.6 | 27.6 | 20.0 | 12.5 | 50.5 |
| Hyperlink Graph | GBS | 20.5 | 13.8 | 72.5 | 17.8 | 14.4 | 23.4 | 23.4 | 15.3 | 49.9 |
| | CGM | 18.4 | 12.0 | 67.0 | 16.4 | 13.2 | 21.5 | 19.4 | 12.7 | 41.3 |
| | HFN | 14.8 | 9.2 | 57.7 | 9.6 | 7.7 | 12.6 | 15.4 | 10.0 | 32.7 |
| Node Embedding | N2V | 17.4 | 11.3 | 63.8 | 15.3 | 12.3 | 20.0 | 18.2 | 11.9 | 38.7 |
| | DW | 16.6 | 10.3 | 63.1 | 11.8 | 9.6 | 15.6 | 16.5 | 10.8 | 35.2 |
| Link Context | BERT | 17.5 | 10.9 | 65.4 | 12.9 | 10.5 | 16.9 | 17.9 | 11.8 | 37.7 |
| | NB | 16.9 | 10.5 | 64.2 | 12.5 | 10.1 | 16.4 | 17.3 | 11.3 | 36.8 |
| | KW | 14.3 | 8.8 | 56.1 | 9.0 | 7.2 | 11.8 | 16.1 | 10.5 | 34.4 |
| | VSM | 13.3 | 8.2 | 54.1 | 7.5 | 6.1 | 9.9 | 12.3 | 8.0 | 26.1 |
| Baseline | BFS | 15.5 | 10.2 | 58.0 | 12.4 | 10.0 | 16.3 | 14.2 | 9.3 | 30.3 |
| | PR | 13.2 | 8.2 | 51.3 | 8.8 | 7.1 | 11.5 | 12.5 | 8.2 | 26.6 |

* All AUC values are calculated as area under the relevant curve across all possible collection cutoffs

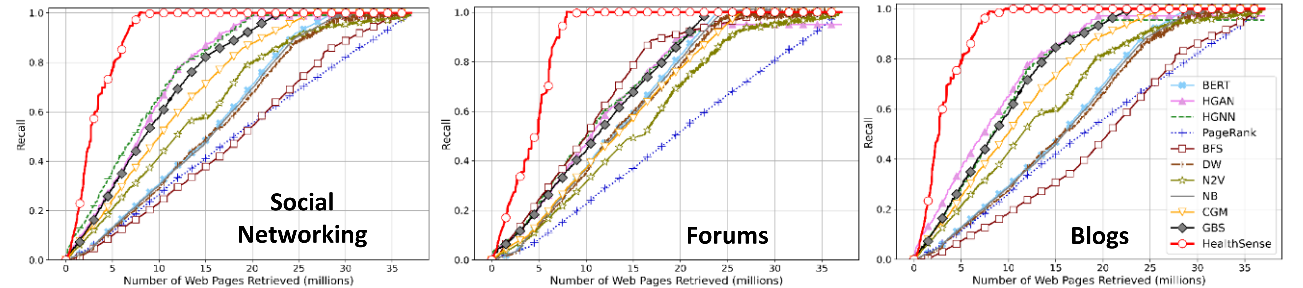


**Figure 7:** Recall Trends across Various Channels (PMDS Task – Opioid Results are Comparable)

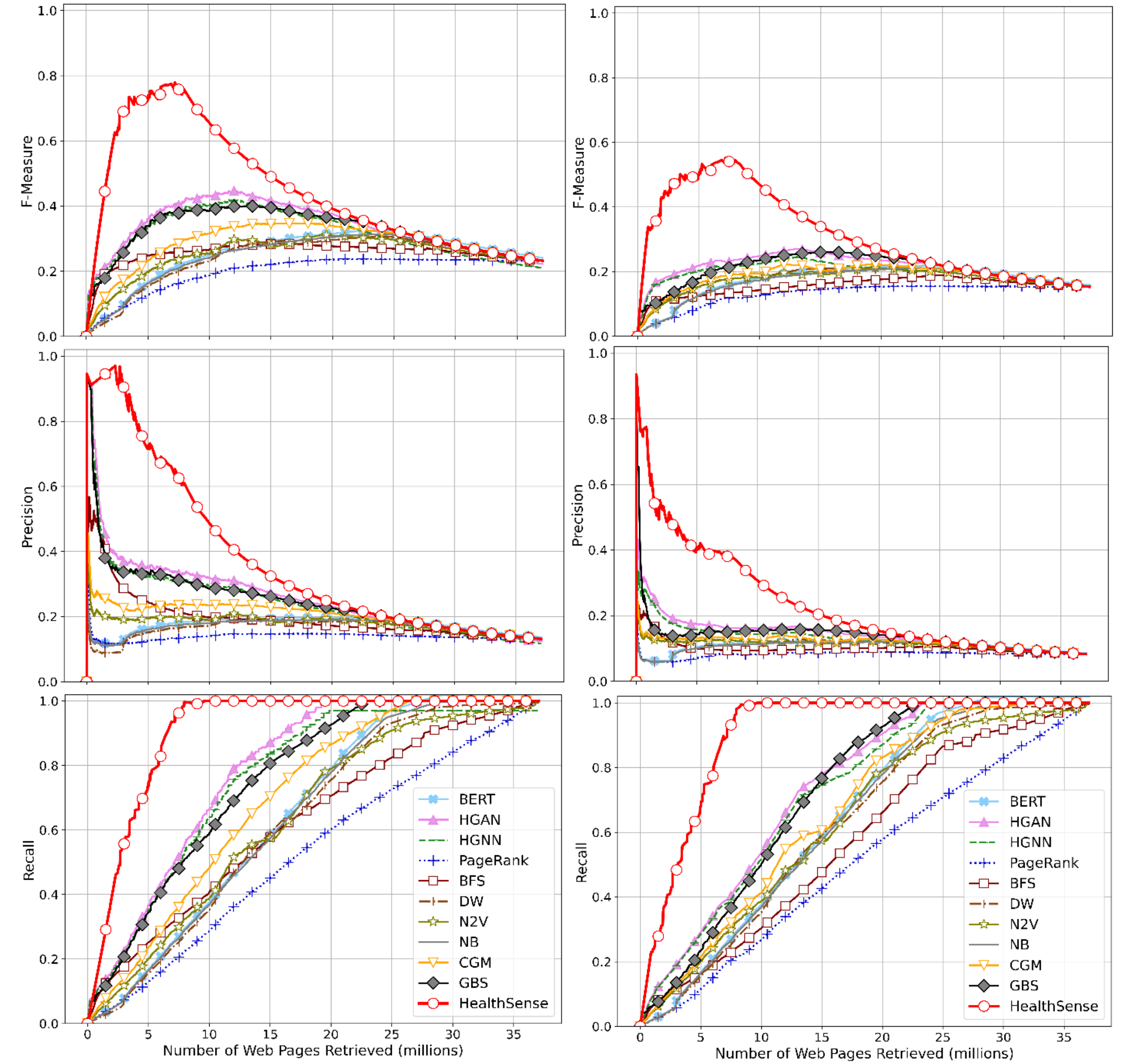
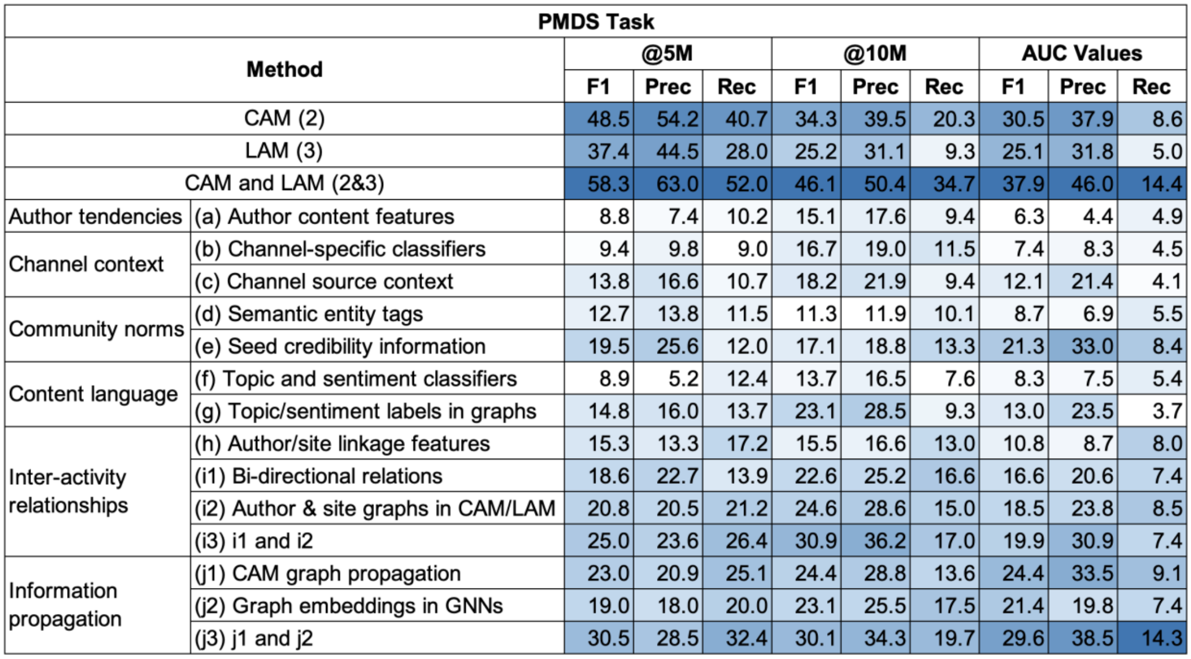


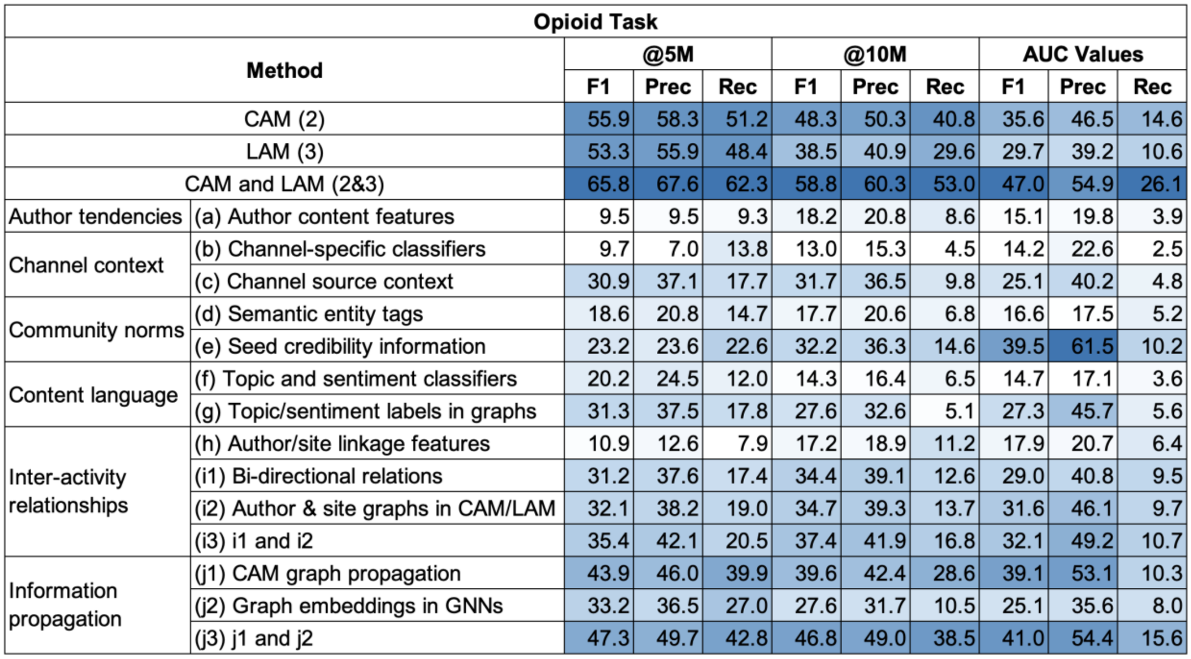
**Figure 8:** Performance Trends for HealthSense & Comparison Methods on PMDS (left) and Opioid (right) Tasks

| **Table 7: Ablation settings** | |
|---|---|
| (a) | Exclude author content features from topic/sentiment classifiers |
| (b) | Replace channel-specific classifiers with a single topic and sentiment classifier |
| (c) | Exclude channel/source features |
| (d) | Omit the lexicon-based semantic entity tags |
| (e) | Replace seed credibility scores with random values between 0-0.1 (similar to PageRank) and set α in eq. 7 to 0 |
| (f) | Exclude topic and sentiment classifiers from RAM, replacing with single relevance classifier per channel |
| (g) | Omit topic and sentiment classification scores from LAM |
| (h) | Exclude author/site linkage features |
| (i1) | Omit bi-directional edge relations from GNNs (eqs. 2-3) and credibility graph (eq. 7) |
| (i2) | Exclude author and site graphs |
| (j1) | Exclude graph propagation, directly using only seed credibility scores in the GNN |
| (j2) | Remove graph embedding (eq. 4), replacing it with a standard dense layer |

**Table 8:** Ablation Analysis: Percent Degradation in HealthSense Performance by Excluded Component



| PMDS Task | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | | **@5M** | | | **@10M** | | | **AUC Values** | | |
| | | **F1** | **Prec** | **Rec** | **F1** | **Prec** | **Rec** | **F1** | **Prec** | **Rec** |
| CAM (2) | | 48.5 | 54.2 | 40.7 | 34.3 | 39.5 | 20.3 | 30.5 | 37.9 | 8.6 |
| LAM (3) | | 37.4 | 44.5 | 28.0 | 25.2 | 31.1 | 9.3 | 25.1 | 31.8 | 5.0 |
| CAM and LAM (2&3) | | 58.3 | 63.0 | 52.0 | 46.1 | 50.4 | 34.7 | 37.9 | 46.0 | 14.4 |
| Author tendencies | (a) Author content features | 8.8 | 7.4 | 10.2 | 15.1 | 17.6 | 9.4 | 6.3 | 4.4 | 4.9 |
| Channel context | (b) Channel-specific classifiers | 9.4 | 9.8 | 9.0 | 16.7 | 19.0 | 11.5 | 7.4 | 8.3 | 4.5 |
| | (c) Channel source context | 13.8 | 16.6 | 10.7 | 18.2 | 21.9 | 9.4 | 12.1 | 21.4 | 4.1 |
| Community norms | (d) Semantic entity tags | 12.7 | 13.8 | 11.5 | 11.3 | 11.9 | 10.1 | 8.7 | 6.9 | 5.5 |
| | (e) Seed credibility information | 19.5 | 25.6 | 12.0 | 17.1 | 18.8 | 13.3 | 21.3 | 33.0 | 8.4 |
| Content language | (f) Topic and sentiment classifiers | 8.9 | 5.2 | 12.4 | 13.7 | 16.5 | 7.6 | 8.3 | 7.5 | 5.4 |
| | (g) Topic/sentiment labels in graphs | 14.8 | 16.0 | 13.7 | 23.1 | 28.5 | 9.3 | 13.0 | 23.5 | 3.7 |
| Inter-activity relationships | (h) Author/site linkage features | 15.3 | 13.3 | 17.2 | 15.5 | 16.6 | 13.0 | 10.8 | 8.7 | 8.0 |
| | (i1) Bi-directional relations | 18.6 | 22.7 | 13.9 | 22.6 | 25.2 | 16.6 | 16.6 | 20.6 | 7.4 |
| | (i2) Author & site graphs in CAM/LAM | 20.8 | 20.5 | 21.2 | 24.6 | 28.6 | 15.0 | 18.5 | 23.8 | 8.5 |
| | (i3) i1 and i2 | 25.0 | 23.6 | 26.4 | 30.9 | 36.2 | 17.0 | 19.9 | 30.9 | 7.4 |
| Information propagation | (j1) CAM graph propagation | 23.0 | 20.9 | 25.1 | 24.4 | 28.8 | 13.6 | 24.4 | 33.5 | 9.1 |
| | (j2) Graph embeddings in GNNs | 19.0 | 18.0 | 20.0 | 23.1 | 25.5 | 17.5 | 21.4 | 19.8 | 7.4 |
| | (j3) j1 and j2 | 30.5 | 28.5 | 32.4 | 30.1 | 34.3 | 19.7 | 29.6 | 38.5 | 14.3 |



| Opioid Task | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | | @5M | | | @10M | | | AUC Values | | |
| | | F1 | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec |
| CAM (2) | | 55.9 | 58.3 | 51.2 | 48.3 | 50.3 | 40.8 | 35.6 | 46.5 | 14.6 |
| LAM (3) | | 53.3 | 55.9 | 48.4 | 38.5 | 40.9 | 29.6 | 29.7 | 39.2 | 10.6 |
| CAM and LAM (2&3) | | 65.8 | 67.6 | 62.3 | 58.8 | 60.3 | 53.0 | 47.0 | 54.9 | 26.1 |
| Author tendencies | (a) Author content features | 9.5 | 9.5 | 9.3 | 18.2 | 20.8 | 8.6 | 15.1 | 19.8 | 3.9 |
| Channel context | (b) Channel-specific classifiers | 9.7 | 7.0 | 13.8 | 13.0 | 15.3 | 4.5 | 14.2 | 22.6 | 2.5 |
| | (c) Channel source context | 30.9 | 37.1 | 17.7 | 31.7 | 36.5 | 9.8 | 25.1 | 40.2 | 4.8 |
| Community norms | (d) Semantic entity tags | 18.6 | 20.8 | 14.7 | 17.7 | 20.6 | 6.8 | 16.6 | 17.5 | 5.2 |
| | (e) Seed credibility information | 23.2 | 23.6 | 22.6 | 32.2 | 36.3 | 14.6 | 39.5 | 61.5 | 10.2 |
| Content language | (f) Topic and sentiment classifiers | 20.2 | 24.5 | 12.0 | 14.3 | 16.4 | 6.5 | 14.7 | 17.1 | 3.6 |
| | (g) Topic/sentiment labels in graphs | 31.3 | 37.5 | 17.8 | 27.6 | 32.6 | 5.1 | 27.3 | 45.7 | 5.6 |
| Inter-activity relationships | (h) Author/site linkage features | 10.9 | 12.6 | 7.9 | 17.2 | 18.9 | 11.2 | 17.9 | 20.7 | 6.4 |
| | (i1) Bi-directional relations | 31.2 | 37.6 | 17.4 | 34.4 | 39.1 | 12.6 | 29.0 | 40.8 | 9.5 |
| | (i2) Author & site graphs in CAM/LAM | 32.1 | 38.2 | 19.0 | 34.7 | 39.3 | 13.7 | 31.6 | 46.1 | 9.7 |
| | (i3) i1 and i2 | 35.4 | 42.1 | 20.5 | 37.4 | 41.9 | 16.8 | 32.1 | 49.2 | 10.7 |
| Information propagation | (j1) CAM graph propagation | 43.9 | 46.0 | 39.9 | 39.6 | 42.4 | 28.6 | 39.1 | 53.1 | 10.3 |
| | (j2) Graph embeddings in GNNs | 33.2 | 36.5 | 27.0 | 27.6 | 31.7 | 10.5 | 25.1 | 35.6 | 8.0 |
| | (j3) j1 and j2 | 47.3 | 49.7 | 42.8 | 46.8 | 49.0 | 38.5 | 41.0 | 54.4 | 15.6 |

* All AUC values are calculated as area under the relevant curve across all possible collection cutoffs

## 6.3 Field Study: HealthSense in Action

The results in the prior section demonstrate the potential for HealthSense to accurately and efficiently retrieve relevant information from a variety of online channels for social listening. To demonstrate the downstream value proposition of these gains, we performed a field study on the PMDS task in conjunction with a major US-based pharmaceutical manufacturer (who, for anonymity, we will refer to as PharmCo). A common use case for safety teams tasked with post-market drug surveillance is to use alternative data sources to examine and verify (or discredit) potential adverse event cases stemming from customer complaints, regulatory queries, or clinical data. Two common methods for examining event cases are (1) manual/qualitative examination of supporting evidence; (2) statistical disproportionality analysis of adverse drug mentions. We used HealthSense to perform field experiments related to both of these case examination methods.

### 6.3.1 Field Experiment Test Bed

To create the test bed for the field experiment, a team of 5 experts from the global drug safety unit at PharmCo used the full PMDS data (encompassing over 37 million data points collected as described previously) to carefully and thoroughly examine 100 cases reported to them over a two-week period. The team used a two-step process to examine the cases. Initially, following standard internal protocols for examining potential adverse drug reaction accounts, team members independently examined the PMDS test bed via custom Tableau dashboards equipped with search capabilities and data visualization, zooming, and filtration functionalities. Each of the five team members individually categorized each case as a “true positive” (case with supporting evidence) or “false positive” (lacking sufficient evidence). The 5 experts then came together to discuss their case assessments and achieve consensus. Ultimately, in their assessment of 100 total cases, the team considered 21 cases to be true positives and the rest to be false positives. This set of cases formed the test bed for both the user experiment and disproportionality analysis described next.

### 6.3.2 User Experiment

In all, 77 members of the global drug safety unit at PharmCo participated in the experiment. None of the 5 experts that assisted with test bed construction participated in the experiment. For purposes of the experiment, 20 cases were selected from the test bed: 10 true positive cases and 10 negatives as determined by the panel of experts. Participants were randomly assigned to one of three experiment groups, each provided with data from a different social listening method: HealthSense, GBS (benchmark), and BFS (baseline). As shown in Table 9, the groups were not significantly different from one another in terms of age, years of experience working on safety/ risk teams, and prior experience working with dashboards (one-way ANOVA p-values > 0.05).

**Table 9:** Summary Statistics for User Study Participants from PharmCo Global Safety Unit

| Group | Age | Safety/Risk Experience | Dashboard Experience |
|---|---|---|---|
| HealthSense Data Users | 36.92 | 10.12 | 3.52 |
| GBS Data Users | 35.77 | 9.95 | 3.61 |
| BFS Data Users | 37.36 | 9.55 | 3.92 |

Within the experiment, all participants used the same Tableau dashboards. Each participant was given 30 minutes of training on how to use the dashboards. They were also provided access to recorded videos with information on how to use various features and functions of the dashboards. For each of the three experimental groups, the @5M post-marketing drug surveillance (PMDS) collection as described in Table 6 was loaded onto the dashboards.

Participants were given 3 hours to examine the 20 cases in the user experiment test bed and categorize each as a "true positive" or "false positive" as previously defined. This duration was chosen to be consistent with the contiguous time blocks that safety team members routinely devote to examining cases. Following the firm's internal protocols and procedures, each participant was also asked to provide written evidence/examples to support their categorizations. The participants' categorizations and supporting responses were examined by the 5 experts. Only those true positive responses with appropriate evidence were considered correct. After the experiment, participants completed a short survey related to the usefulness of the PMDS data and dashboards provided. The results are summarized in Table 10. Participants using the HealthSense data were not only able to identify true positives with significantly higher precision and recall than those using GBS or BFS data, but were also significantly better at identifying false positive cases. These results suggest HealthSense effectively supports the societal benefits of faster, more effective adverse event identification while simultaneously reducing PMDS investigation costs.

**Table 10:** Field Experiment Results for HealthSense and Comparison Methods

| Group | PTP Precision | PTP Recall | PTP F-Measure | FP Precision | FP Recall | FP F-Measure |
|---|---|---|---|---|---|---|
| HealthSense Data Users | **85.34** | **81.54** | **82.97** | **83.31** | **85.77** | **84.12** |
| GBS Data Users | 70.24 | 65.00 | 65.99 | 69.05 | 71.15 | 68.87 |
| BFS Data Users | 63.32 | 63.08 | 62.25 | 66.58 | 66.54 | 65.50 |

### 6.3.3 Disproportionality Analysis Case Study

In addition to its usefulness for vetting reported adverse drug events, many stakeholders, including pharmaceutical companies, regulators, and healthcare hedge funds, stand to benefit from early detection of such adverse events prior to reporting. Disproportionality analysis (Rothman et al. 2004) is a commonly used technique for automatically detecting adverse events from various data sources by comparing the occurrences and co-occurences of entities and outcomes in a corpus. To see how accurately these events could be detected solely based on data gathered by our social listening platform as compared to others, we performed disproportionality analysis using data from HealthSense, GBS, and BFS at the 5 million URL collection threshold.

For purposes of the disproportionality analysis, we measured co-occurrence of drug and reaction tuples within documents using the reporting odds ratio (ROR) metric. Tuples with 95% confidence of ROR≥1.0 were considered positive predictions. Note that since disproportionality analysis was performed at the drug-reaction tuple level, cases comprised of more than one reaction related to a drug could allow for multiple drug-reaction true positives for the same case. Performance was evaluated using recall of the 21 possible true positive cases, and precision defined as proportion of all positive signals that related to true positives identified by the experts. Table 11 summarizes results. Relative to the user experiment, HealthSense demonstrated even higher performance gains. Using HealthSense data, recall for the disproportionality analysis was 36% higher and precision 81% higher than that achieved using GBS. Through the use of data efficiently gathered by HealthSense, it is clear that pharmaceutical companies or other stakeholders could effectively use social listening to discover significantly more true adverse events, while substantially reducing time and resources spent investigating false positives.

**Table 11:** Event Detection Results for HealthSense and Comparison Methods

| ROR Data | Unique Event Cases Detected | True Positive Signals | False Positive Signals | Case Recall | Signal Precision |
|---|---|---|---|---|---|
| HealthSense Data | **15** | **23** | **26** | **71.43** | **46.94** |
| GBS Data | 11 | 14 | 40 | 52.38 | 25.93 |
| BFS Data | 10 | 12 | 42 | 47.62 | 22.22 |

# 7 Discussion and Conclusion

In this study, we propose HealthSense – an effective, efficient social listening platform that can be used to provide timely, granular, and actionable data for time-sensitive analysis in support of public health tasks. Through a variety of real-world use-cases, we demonstrate that it is capable of providing significant value and improvements in public health outcomes, especially in time-sensitive situations. It significantly outperforms currently available tools for social listening – creating new possibilities for aiding time-sensitive public health informatics.

From a computational design perspective (Rai 2017; Padmanabhan et al. 2022), we make five distinct contributions to research and practice. First, we show that theory-guided social listening artifacts can seamlessly combine relevance, credibility, and cross-channel landscape assessment to enable markedly better public health sensing capabilities. In order to meet the challenges of the 21st century, including improving the social determinants of health (Wang and DeSalvo 2018), social listening platforms must shift from the data collection and information retrieval paradigm towards intelligent sensing. In the same vein as prior IS design research in societally impactful contexts such as emergency response (Chen et al. 2013), our work underscores the value of using Activity Theory to understand the complexities and nuances of online social activities and content creation via digital communication. The combination of the Activity Theory framework along with theories regarding contextual information quality were instrumental in facilitating more accurate and efficient listening capabilities in our proposed artifact.

Second, through our design process, we develop and propose an extension to existing Activity Theory literature, which addresses gaps in its ability to describe communication activities on modern online platforms – specifically the ability to capture information hidden in the multiplex relationships between communication activities. As part of our design framework developed based on this extension, we propose novel credibility and relevance assessment mechanisms that use graph propagation and content-based classifiers in conjunction with state-of-the-art graph convolutional networks that employ node and edge-conditioned embeddings. We then use rigorous and detailed evaluation of our artifact to empirically demonstrate that the inclusion of these mechanisms based on our extension to Activity Theory drastically improves the capabilities of our proposed social listening artifact. This novel contribution back to theory and demonstration of its value opens new opportunities for researchers to use the Activity Theoretic lens to address further issues relating to communication activities on modern online platforms.

Third, in designing our artifact, we contribute to knowledge on graph neural networks through two novel extensions to this methodology. First, to address issues related to label sparsity, we couple graph propagation methods with graph neural networks. This allows us to propagate sparsely known seeding information across the network, improving performance of GNNs which perform better in low-sparsity environments. This extends limited prior work in this area by Bojchevski et al. (2020). Second, we extend the GNN architecture to include bi-relational edge-enriched node embeddings informed by domain-adapted feature-based classifiers. This allows for parsimonious representation of crucial information about in- and out-bound links to be utilized in relevance and landscape assessment. Through our ablation analysis, both of these novel extensions were shown to add significant value to our social listening artifact.

Fourth, we show that in dynamic environments involving machine learning applied to complex user-generated content, artifacts guided by human-centered theories and intuition remain critical complements to automated AI-driven techniques. State-of-the-art learning representations such as graph convolutional networks (Wu et al. 2021) using edge-conditioned node embeddings represent powerful new methods for deriving graph-based patterns related to credibility and relevance. However, in emergent contexts such as the early stages of social listening, these methods are at their best when used in combination with theory-supported methods to represent heterogeneity and extract context from limited available information. We believe our work is a microcosm of how technical research can combine cumulative knowledge with state-of-the-art machine learning methods in the era of large-scale pre-trained embeddings and multi-billion parameter universal language models.

Finally, our results across data, user, and event experiments demonstrate the downstream value chain associated with effectively designed social listening artifacts. Calls from champions of the public health 3.0 movement for timely, granular, and actionable data point directly to the prospects of tangible value that may be obtained through improved public health outcomes (Wang and DeSalvo 2018). Yet, without direct evidence, the presumption of value from such data and related analyses remains unclear. In our study, we show that HealthSense identified over 90% of relevant information for specified tasks by analyzing less than 20% of data. Through our partnership with the pharmacovigilance team at a major US pharmaceutical manufacturer, we further are able to show how this improvement translates into actual improvements in the downstream value chain. Data gathered by HealthSense resulted in a 36% improvement in recall and 81% improvement in precision in automated analysis, and a 22% improvement in recall and 25% improvement in precision for downstream manual investigation of potential adverse drug reactions. These real-world improvements in both effectiveness and efficiency of downstream analytics point directly to the impacts on public health outcomes made possible by HealthSense.

Our work also has potential applications beyond those in the public health domain. Social listening capabilities can be beneficial in a variety of contexts—a prime example is digital marketing. Although social listening is prevalent in this context, marketing research and practice have largely focused on use cases that do not require near-real-time analytics, such as utilizing user generated content to complement or replace traditional market research (Tirunillai and Tellis 2014). Alternatively, when near-real-time analysis is an important aspect of the use case, the focus is often on a single platform with ready access to data through native tools or APIs, such as in the case of targeted advertising and messaging based on social media activity (Adamopoulos et al. 2018). Our system has significant utility specifically in circumstances where time is of the essence and information is required from a wide variety of online channels. For instance, in identifying and responding to users omnichannel product discussions and purchase behaviors (Cui et al. 2021; Sun et al. 2022), or in crisis identification and response, detecting groundswells of negative reaction to products or company activities (Hewett et al. 2016). Work in these areas has largely, to date, been focused on retrospective analysis, but in practice would require a social listening tool such as ours in order to enable near-real-time responses.

Our work is not without its limitations. Importantly, our models are only able to account for a small portion of the totality of complexity present in online communications. For instance, beyond establishing norms that contextualize communications, online communities provide social structures within which authors and their contributions are evaluated, which provides further context. Complexities such as this are difficult to capture, but could provide significant value. Our artifact also requires startup costs in the form of the creation of small task-relevant training sets for seeding. So, while useful for monitoring time-sensitive information for stable tasks, it may be less effective for exploratory tasks devoid of domain knowledge. Future extensions could address this issue. Our listening testbeds also focused on text-based content. Recent work has underscored the importance of health-related video and audio content (Li et al. 2019, Liu et al. 2020), and we believe future work should extend public health listening to multimedia contexts. In sum, despite these acknowledged limitations, we believe this work has important implications for IS research at the intersection of design and data science that integrates social-technical concepts into novel domain-adapted machine learning artifacts in societally impactful contexts, and for practitioners requiring data from online platforms to fuel time-sensitive informatics.

## Acknowledgements

This work was funded in part through U.S. NSF grant IIS-2039915.

## References

Abbasi, A., Zahedi, F. M., Kaza, S. 2012. "Detecting Fake Medical Web Sites Using Recursive Trust Labeling," ACM Transactions on Information Systems (30:4), p. 22.

Abbasi, A., Zhou, Y., Deng, S., Zhang, P. 2018. "Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective," MIS Quarterly (42:2), pp. 427–464.

Adamopoulos, P., Ghose, A., Todri, V. 2018. "The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms," ISR (29:3), pp. 612–640.

Adjeroh, D., Beal, R., Abbasi, A., Zheng, W., Abate, M., Ross, A. 2014. "Signal Fusion for Social Media Analysis of Adverse Drug Events," IEEE Intelligent Sys (29:2), pp. 74–80.

Aggarwal, C. C., Al-Garawi, F., Yu, P. S. 2001. "Intelligent Crawling on the World Wide Web with Arbitrary Predicates," International Conference on WWW, NY, USA.

Allen, D. K., Brown, A., Karanasios, S., Norman, A. 2013. "How Should Technology-Mediated Organizational Change Be Explained? A Comparison of the Contributions of Critical Realism and Activity Theory," MIS Quarterly (37:3), pp. 835–854.

AP. 2016. "Drug Dealers Are Mixing Heroin with Elephant Sedatives," Associated Press.

Apple Media. "Apple and Google Partner on COVID-19 Contact Tracing Technology." 2020. https://www.apple.com/newsroom/2020/04/apple-and-google-partner-on-covid-19-contact-tracing-technology/.

Arazy, O., Kopak, R. (2011). On the measurability of information quality. JASIST, 62(1), 89–99.

Arazy, O., Kopak, R., & Hadar, I. 2017. "Heuristic Principles and Differential Judgments in the Assessment of Information Quality," JAIS, 18(5), 403–432.

Arazy, O., Kumar, N., & Shapira, B. 2010. "A theory-driven design framework for social recommender systems," JAIS, 11(9).

Baccianella, S., Esuli, A., Sebastiani, F. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," LREC, Valletta, Malta.

Barnes, S. S., Kaul, V., Kudchadkar, S. R. 2019. "Social Media Engagement and the Critical Care Medicine Community," Journal of Intensive Care Medicine (34:3), pp. 175–182.

Benamar, L., Balagué, C., Ghassany, M. 2017. "The Identification and Influence of Social Roles in a Social Media Product Community," JCMC (22:6), pp. 337–362.

Bender, J. L., O'Grady, L., Jadad, A. R. 2008. "Supporting Cancer Patients through the Continuum of Care: A View from the Age of Social Networks and Computer-Mediated Communication," Current Oncology (15), pp. S42–S47.

Bergmark, D., Lagoze, C., Sbityakov, A. 2002. "Focused Crawls, Tunneling, and Digital Libraries," European Conference on Digital Libraries, London, UK.

Bojchevski, A., Klicpera, J., Perozzi, B., Kapoor, A., Blais, M., Rózemberczki, B., Lukasik, M., Günnemann, S. 2020. "Scaling Graph Neural Networks with Approximate PageRank," Intl Conference on Knowledge Discovery & Data Mining, CA, USA.

Boldi, P., Marino, A., Santini, M., Vigna, S. 2018. "BUbiNG: Massive Crawling for the Masses," ACM Transactions on the Web (12:2), 12:1-12:26.

Boudry, C. 2015. “Web 2.0 Applications in Medicine: Trends and Topics in the Literature,” Medicine 2.0 (4:1), p. e2.

Bowen, D. A., O’Donnell, J., Sumner, S. A. 2019. “Increases in Online Posts About Synthetic Opioids Preceding Increases in Synthetic Opioid Death Rates: A Retrospective Observational Study,” Journal of General Internal Medicine (34:12), pp. 2702–2704.

Brewer, T., Colditz, G. A. 1999. “Postmarketing Surveillance and Adverse Drug Reactions: Current Perspectives and Future Needs,” JAMA (281:9), pp. 824–829.

Brin, S., Page, L. 1998. “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Computer Networks and ISDN Systems (30:1–7), pp. 107–117.

Broniatowski, D. A., Paul, M. J., Dredze, M. 2014. “Twitter: Big Data Opportunities,” Science (New York, N.Y.) (345:6193), p. 148.

Cameron, D., Smith, G. A., Daniulaityte, R., Sheth, A. P., Dave, D., Chen, L., Anand, G., Carlson, R., Watkins, K. Z., & Falck, R. 2013. “PREDOSE: A semantic web platform for drug abuse epidemiology using social media,” Journal of Biomedical Informatics, 46(6), 985–997.

CDC. “Social Listening and Monitoring Tools.” 2021. https://www.cdc.gov/vaccines/covid-19/vaccinate-with-confidence/rca-guide/downloads/cdc_rca_guide_2021_tools_appendixe_sociallistening-monitoring-tools-508.pdf.

Chau, M., Chen, H. 2003. Comparison of Three Vertical Search Spiders.

Chau, M., Chen, H. 2007. “Incorporating Web Analysis into Neural Networks: An Example in Hopfield Net Searching,” IEEE TSMC (37:3), pp. 352–358.

Chen, H. C., Lally, A. M., Zhu, B., Chau, M. 2003. “HelpfulMed: Intelligent Searching for Medical Information over the Internet,” JASIST (54:7), pp. 683–694.

Chen, R., Rao, H. R., Sharman, R., Upadhyaya, S. J., Chakravarti, N. 2008. “Emergency Response Information System Interoperability: Development of Chemical Incident Response Data Model,” JAIS (9:3–4), pp. 200–230.

Chen, R., Sharman, R., Rao, H. and Upadhyaya, S. 2013. “Data Model Development for Fire Related Extreme Events: An Activity Theory Approach,” MIS Quarterly.(37:1),pp. 125–147

Cho, J., Garcia-Molina, H., Page, L. 1998. “Efficient Crawling Through URL Ordering,” International Conference on WWW, Amsterdam, The Netherlands, pp. 161–172.

Chung, W., Chen, H., Nunamaker, J. F. 2005. “A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration,” JMIS (21:4), pp. 57–84.

Cole-Lewis, H., Perotte, A., Galica, K., Dreyer, L., Griffith, C., Schwarz, M., Yun, C., Patrick, H., Coa, K., Augustson, E. 2016. “Social Network Behavior and Engagement Within a Smoking Cessation Facebook Page,” JMIR (18:8), pp. 187–197.

Cui, T. H., Ghose, A., Halaburda, H., Iyengar, R., Pauwels, K., Sriram, S., Tucker, C., Venkataraman, S. 2021. “Informational Challenges in Omnichannel Marketing: Remedies and Future Research,” Journal of Marketing (85:1), pp. 103–120.

Dadgar, M., Joshi, K. D. 2018. “The Role of Information and Communication Technology in Self-Management of Chronic Diseases: An Empirical Investigation through Value Sensitive Design,” JAIS (19:2), pp. 86–112.

Dai, E., Jin, W., Liu, H., Wang, S. 2022. “Towards Robust Graph Neural Networks for Noisy Graphs with Sparse Labels,” Intl Conf on Web Search and Data Mining, NY, USA.

Davis, M., Logan, D. 2019. “Market Guide for Social Analytics Applications,” Gartner.

Davison, B. D. 2000. “Topical Locality in the Web,” Conference on Research and Development in Information Retrieval, New York, NY, USA.

DeMio, T. 2016. “Warning: Opioid for Elephants Hitting Ohio Streets,” USA Today Network.

DeSalvo, K. B., Wang, Y. C., Harris, A., Auerbach, J., Koo, D., O’Carroll, P. 2017. “Public Health 3.0: A Call to Action for Public Health to Meet the Challenges of the 21st Century,” Preventing Chronic Disease (14).

Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., Gori, M. 2000. “Focused Crawling Using Context Graphs,” Intl Conf on Very Large Data Bases, San Francisco, CA, USA.

Effah, J., Adam, I. 2021. “Examining Client-Vendor Relationship in the Outsourcing of a Work Environment Virtualisation: An Activity Theory Perspective,” ISF (24:1).

Engeström. 1987. Learning by Expanding. An Activity-Theoretical Approach to Developmental Research, Helsinki, Finland.: Orienta-Konsultit Oy.

Farag, M., Lee, S., Fox, E. A. 2018. “Focused Crawler for Events,” Intl J on Digital Libraries (19:1), pp. 3–19.

Fichman, R. G., Kohli, R., Krishnan, R. 2011. “The Role of Information Systems in Healthcare: Current Research and Future Trends,” ISR (22:3), pp. 419–428.

Fogg, B. J. 2003. “Prominence-interpretation theory: Explaining how people assess credibility online,” CHI ’03 Extended Abstracts on Human Factors in Computing Systems, 722–723.

Foley, S., Karlsen, J. R., Putniņš, T. J. 2018. “Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed Through Cryptocurrencies?,” SSRN Scholarly Paper No. ID 3102645.

Fu, T., Abbasi, A., Zeng, D., Chen, H. 2012. “Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers,” ACM Transactions on IS (30:4), p. 24.

Garrett, B., Murphy, S., Jamal, S., MacPhee, M., Reardon, J., Cheung, W., Mallia, E., Jackson, C. 2019. “Internet Health Scams-Developing a Taxonomy and Risk-of-Deception Assessment Tool,” Health & Social Care in the Community (27:1), pp. 226–240.

Ge, M., Helfert, M., Jannach, D. 2011. “Information Quality Assessment: Validating Measurement Dimensions and Processes,” ECIS, Helsinki, Finland.

van Grootheest, K., de Graaf, L., de Jong-van Den Berg, L. 2003. “Consumer Adverse Drug Reaction Reporting: A New Step in Pharmacovigilance?,” Drug Safety (26:4), pp. 211–217.

Grover, A., Leskovec, J. 2016. “Node2vec: Scalable Feature Learning for Networks,” in Intl Conf on Knowledge Discovery and Data Mining, San Francisco, California USA.

Gyöngyi, Z., Garcia-Molina, H., Pedersen, J. 2004. “Combating Web Spam with Trustrank,” Intl Conf on Very Large Data Bases, Toronto, Canada.

Hamilton, W., Ying, Z., Leskovec, J. 2017. “Inductive Representation Learning on Large Graphs,” in Advances in Neural Info Processing Systems (Vol. 30), Curran Associates, Inc.

Hansen, M. M. 2008. “Versatile, Immersive, Creative and Dynamic Virtual 3-D Healthcare Learning Environments: A Review of the

Literature," JMIR (10:3), p. e26.

Hayes, D. R., Cappa, F., Cardon, J. 2018. "A Framework for More Effective Dark Web Marketplace Investigations," Information (9:8), p. 186.

Hewett, K., Rand, W., Rust, R. T., van Heerde, H. J. 2016. "Brand Buzz in the Echoverse," Journal of Marketing (80:3), pp. 1–24.

Joachims, T. 2006. "Training Linear SVMs in Linear Time," Intl Conf on Knowledge Discovery and Data Mining, Philadelphia, PA, USA.

Kaptelinin, V., Nardi, B. 2006. Acting with Technology: Activity Theory and Interaction Design, Cambridge, MA: MIT Press.

Kipf, T. N., Welling, M. 2017. "Semi-Supervised Classification with Graph Convolutional Networks," ArXiv:1609.02907 [Cs, Stat].

Kitchin, R., McArdle, G. 2016. "What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets," Big Data & Society (3:1).

Knight, S., Burn, J. 2005. "Developing a Framework for Assessing Information Quality on the World Wide Web," Informing Science Journal (8), pp. 159–173.

Kobayashi, M., Takeda, K. 2000. "Information Retrieval on the Web," ACM Computing Surveys (32:2), pp. 144–173.

Korpela, M., Mursu, A., Soriyan, H. A. 2002. "Information Systems Development as an Activity," Computer Supported Cooperative Work (11:1–2), pp. 111–128.

Kumar, M., Bhatia, R., Rattan, D. 2017. "A Survey of Web Crawlers for Information Retrieval," WIREs Data Mining and Knowledge Discovery (7:6), p. e1218.

Kuutti, K. 1991. "Activity Theory and Its Applications to Information Systems Research and Development," ISR, pp. 529–549.

Ladegaard, I. 2019. "Crime Displacement in Digital Drug Markets," Int J of Drug Policy (63).

Lazer, D., Kennedy, R., King, G., Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis," Science (343:6176), pp. 1203–1205.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G. 2010. "Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks," Workshop on Biomedical NLP, 117–125.

Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y. 2002. "AIMQ: A Methodology for Information Quality Assessment," Information & Management (40:2), pp. 133–146.

Leontev, A. N. 1978. Activity, Consciousness, and Personality. Prentice-Hall.

Li, M., Yan, S., Yang, D., Li, B., Cui, W. 2019. "YouTube (TM) as a Source of Information on Food Poisoning," BMC Public Health (19), p. 952.

Liu, X., Zhang, B., Susarla, A., Padman, R. 2020. "Go to YouTube and Call Me in the Morning: Use of Social Media for Chronic Conditions," MIS Quarterly (44:1b), pp. 257–283.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit," Association for Computational Linguistics.

McClellan, C., Ali, M., Mutter, R., Kroutil, L., Landwehr, J. 2017. "Using Social Media to Monitor Mental Health Discussions - Evidence from Twitter," JAMIA (24:3), pp. 496–502.

Menczer, F., Pant, G., Srinivasan, P. 2004. "Topical Web Crawlers: Evaluating Adaptive Algorithms," ACM Trans. Internet Technol. (4:4), pp. 378–419.

Nasralah, T., El-Gayar, O., & Wang, Y. 2020. Social Media Text Mining Framework for Drug Abuse: Development and Validation Study with an Opioid Crisis Case Analysis. JMIR, 22(8).

Nelson, R. R., Todd, P. A., Wixom, B. H. 2005. "Antecedents of Information and System Quality," JMIS (21:4), pp. 199–235.

Overbeek, D., Janke, A. 2018. "360 Characteristics of Posts of Opioid Users on Reddit, an Online Social Media Forum, an Area for Improved Harm Reduction," Annals of Emer Med.

Padmanabhan, B., Sahoo, N., Burton-Jones, A. 2022. "Machine Learning in Information Systems Research," MIS Quarterly (46:1), pp. iii–xix.

Pagoto, S., Waring, M. E., Xu, R. 2019. "A Call for a Public Health Agenda for Social Media Research," JMIR (21:12).

Pandrekar, S., Chen, X., Gopalkrishna, G., Srivastava, A., Saltz, M., Saltz, J., Wang, F. 2018. "Social Media Based Analysis of Opioid Epidemic Using Reddit," AMIA Symposium.

Pant, G., Srinivasan, P. 2005. "Learning to Crawl: Comparing Classification Schemes," ACM Transactions on IS (23:4), pp. 430–462.

Park, A., Conway, M., Chen, A. T. 2018. "Examining Thematic Similarity, Difference, and Membership in Three Online Mental Health Communities from Reddit: A Text Mining and Visualization Approach," Computers in Human Behavior (78), pp. 98–112.

Peacock, S., Reddy, A., Leveille, S. G., Walker, J., Payne, T. H., Oster, N. V., Elmore, J. G. 2017. "Patient Portals and Personal Health Information Online: Perception, Access, and Use by US Adults," JAMIA (24: 1), pp. e173–e177.

Perozzi, B., Al-Rfou, R., Skiena, S. 2014. "DeepWalk: Online Learning of Social Representations," Intl Conf on Knowledge Discovery and Data Mining, New York, USA.

Pour, M. J., Jafari, S. 2018. "Toward a Maturity Model for the Application of Social Media in Healthcare: The Health 2.0 Roadmap," Online Information Review (43).

Rai, A. 2017. "Editor's comments: Diversity of Design Science Research," MIS Quarterly, 41(1).

Riloff, E., Patwardhan, S., Wiebe, J. 2006. "Feature Subsumption for Opinion Analysis," Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.

Rothman, K., Lanes, S., Sacks, S. 2004. "The Reporting Odds Ratio and Its Advantages over the Proportional Reporting Ratio," Pharmacoepidemiology Drug Safety(13:8), pp. 519–523.

Saloner, B., Chang, H., Ferris, L., Eisenberg, M., Richards, T., Lemke, K., Schneider, K., Baier, M., Weiner, J. 2020. "Predictive Modeling of Opioid Overdose Using Linked Statewide Medical and Criminal Justice Data," JAMA Psych(77:11), pp. 1155–1162.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G. 2015. "Utilizing Social Media Data for Pharmacovigilance: A Review," Journal of Biomedical Informatics (54), pp. 202–212.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. van den, Titov, I., Welling, M. 2017. "Modeling Relational Data with Graph Convolutional Networks," ArXiv:1703.06103.

Schwetz, T. A., Calder, T., Rosenthal, E., Kattakuzhy, S., Fauci, A. S. 2019. "Opioids and Infectious Diseases: A Converging Public Health Crisis," The J of Infectious Diseases.

Shoff, E. N., Zaney, M. E., Kahl, J. H., Hime, G. W., Boland, D. M. 2017. "Qualitative Identification of Fentanyl Analogs and Other Opioids in Postmortem Cases by UHPLC-Ion Trap-MSn," Journal of Analytical Toxicology (41:6), pp. 484–492.

Smith, A. N., Fischer, E., Yongjian, C. 2012. "How Does Brand-Related User-Generated Content Differ across YouTube, Facebook, and

Twitter?," JIM (26:2), pp. 102–113.

Song, J., Zahedi, F. M. 2007. "Trust in Health Infomediaries," DSS (43:2), pp. 390–407.

Squirrell, T. 2019. "Platform Dialectics: The Relationships between Volunteer Moderators and End Users on Reddit," New Media & Society (21:9), pp. 1910–1927.

Srinivasan, P., Menczer, F., Pant, G. 2005. "A General Evaluation Framework for Topical Crawlers," Information Retrieval (8:3), pp. 417–447.

Strong, D. M., Lee, Y. W., Wang, R. Y. 1997. Data Quality in Context. CACM, 40(5), 103–110.

Sun, C., Adamopoulos, P., Ghose, A., Luo, X. 2022. "Predicting Stages in Omnichannel Path to Purchase: A Deep Learning Model," ISR (33:2), pp. 429–445.

Szalavitz, M. 2011. "Mayo Clinic vs. WebMD: Another Perspective," Time.

Tirunillai, S., & Tellis, G. J. 2014. "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," JMR, 51(4), 463–479.

Todoran, I., Lecornu, L., Khenchaf, A., Caillec, J. 2015. "A Methodology to Evaluate Important Dimensions of Information Quality in Systems," JDIQ (6:23), 1-23.

Valecha, R., Rao, R., Upadhyaya, S., Sharman, R. 2019. "An Activity Theory Approach to Modeling Dispatch-Mediated Emergency Response," JAIS (20:1).

Veličković, P., Casanova, A., Lio, P., Cucurull, G., Romero, A., Bengio, Y. 2018. "Graph Attention Networks," Conference on Learning Representations, San Juan, Puerto Rico.

Viviani, M., Pasi, G. 2017. "Credibility in Social Media: Opinions, News, and Health Information—a Survey," WIREs Data Mining and Knowledge Discovery (7:5), p. e1209.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. 1992. "Building an information system design theory for vigilant EIS," ISR 3(1), 36-59.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Yu, P. S. 2019. "Heterogeneous Graph Attention Network," The World Wide Web Conference, San Francisco, CA.

Wang, Y. C., DeSalvo, K. 2018. "Timely, Granular, and Actionable: Informatics in the Public Health 3.0 Era," American Journal of Public Health (108:7), pp. 930–934.

Waterloo, S. F., Baumgartner, S. E., Peter, J., Valkenburg, P. M. 2018. "Norms of Online Expressions of Emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp," New Media & Society (20:5), pp. 1813–1831.

Weng, J., Lim, E.-P., Jiang, J., He, Q. 2010. "TwitterRank: Finding Topic-Sensitive Influential Twitterers," ACM Int Conf on Web Search and Data Mining, New York, NY.

Wiebe, J., Wilson, T., & Cardie, C. 2005. "Annotating Expressions of Opinions and Emotions in Language," LREC, 39(2), 165–210.

Wilson, N. 2020. "Drug and Opioid-Involved Overdose Deaths — United States, 2017–2018," MMWR. Morbidity and Mortality Weekly Report (69).

Wright, A., Bates, D. W., Middleton, B., Hongsermeier, T., Kashyap, V., Thomas, S. M., Sittig, D. F. 2009. "Creating and Sharing Clinical Decision Support Content with Web 2.0: Issues and Examples," Journal of Biomedical Informatics (42:2), pp. 334–346.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P. 2021. "A Comprehensive Survey on Graph Neural Networks," IEEE Trans on Neural Nets \ Learning Systems (32:1), pp. 4–24.

Yan, L., Tan, Y. 2014. "Feeling Blue? Go Online: An Empirical Study of Social Support Among Patients," ISR (25:4), pp. 690–709.

Yang, K., Lau, R. Y. K., Abbasi, A. 2023. "Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality," Information Systems Research, 34(1), 194-222.

Yang, Y. T., Horneffer, M., DiLisio, N. 2013. "Mining Social Media and Web Searches for Disease Detection," Journal of Public Health Research (2:1), pp. 17–21.

Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N. V. 2019. "Heterogeneous Graph Neural Network," Intl Conf on Knowledge Discovery & Data Mining, Anchorage, AK.

Zheng, Yiming, Zhao, K., Stylianou, A. 2013. "The Impacts of Information Quality and System Quality on Users' Continuance Intention in Information-Exchange Virtual Communities: An Empirical Investigation," DSS (56), pp. 513–524.

Zimbra, D., Abbasi, A., Zeng, D., Chen, H. 2018. "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation," ACM TMIS (9:2), 5:1-5:29.