Editors: **Ahmed Abbasi,** University of Virginia, abbasi@comm.virginia.edu
Donald Adjeroh, West Virginia University, don@csee.wvu.edu

Social Media Analytics for Smart Health

Ahmed Abbasi, *University of Virginia*
Donald Adjeroh, *West Virginia University*

In recent years, there has been a growing emphasis on the development of information and communication technologies capable of providing users with mechanisms for generating and accessing medical content via social media. Consequently, people are increasingly turning to social media as a source for health-related information and products, to seek information about health and wellness, and to attain advice, share experiences, and voice concerns. Furthermore, social media, virtual worlds, and mobile applications are increasingly being used to provide new avenues for healthcare delivery. The richness and abundance of health-related social media content has opened up an array of analytical possibilities.

Understanding the importance of various channels in the context of smart health and real-time analytics is an important, yet little-explored endeavor. Many stakeholder groups, including patients, physicians, hospitals, pharmaceutical companies, and regulatory agencies are interested in knowing which channels are most suitable with respect to various dimensions. Meta-analysis of social media is warranted to provide empirical insights regarding the various channels' *credibility*, *recency*, *uniqueness*, *frequency*, and *salience* (CRUFS), as Figure 1 shows.

Credibility refers to the quality, trustworthiness, and integrity of information appearing in social media. Given that social media revolves around user-generated content, the credibility of the source and message are critical. Recency pertains to the timeliness of the information and insights attained via social media. For some tasks, social media is considered a potential early-warning indicator. For others, it is more suitable as a proxy for user responses and reactions. Furthermore, these results may be very channel-dependent. A central tenet of social media is the ability to share and propagate news, opinions, and ideas across channels. Consequently, the uniqueness of information for a particular task may vary across channels in the era of the viral Web, where message board comments with links to news articles and re-tweets are the norm. Although on one hand this phenomenon is analogous to crowdsourcing information importance, on the other, it introduces duplicity and possible redundancy, since certain sources are being amplified. Frequency is an interesting challenge with social media, because some channels clearly have larger volumes than others.[3] However, volume is also often inversely proportional to salience, including the extent to which health topics are discussed, the manner in which ideas are articulated, the depth of opinions and emotions expressed, and the richness of narratives. Although greater *salience*—that is, the ability to perform deeper natural language processing (NLP) to understand who, what, when, where, why, and how—is welcomed generally for health-related analytical tasks, such salience also raises privacy concerns. At what point does usage of open social media data and public forums become invasive?

Here, we briefly elaborate on the five facets of the CRUFS framework and how each has important implications for social media analytics for smart health. We then introduce the articles appearing in this "Trends & Controversies" special section, which touch upon many interesting aspects of CRUFS.

Credibility

Prior studies have examined credibility levels in certain channels. For instance, between 4 and 10 percent of reviews posted on four websites examined were spam.[1] Similarly, more than 10 percent of blogs are estimated to be spam. It has been estimated that up to 20 percent of all pages on the Web are spam, while more than 40 percent of all pages examined on the .US and .Biz domains were not credible.[2]

1541-1672/14/$31.00 © 2014 IEEE
Published by the IEEE Computer Society

A similar credibility assessment of various channels in the context of health-related information is warranted. As Figure 2 shows, we analyzed thousands of drug and health-related websites known to be credible/non-credible based on data taken from third-party health URL databases such as LegitScript, the National Association of Boards of Pharmacy (NABP), Health on Net, and the Consumer and Patient Health Information Section (CAPHIS 100).[2,3] The hyperlink graph depicts site nodes colored based on credibility (gray nodes' credibility status was unknown). Interestingly, the credible sites are closely connected, while less-credible content is dispersed across various network cliques. Figure 2 underscores the prevalence and reach of non-credible medical content across the Web. In the context of social media analytics for smart health, examination of the nature, motives, and sources of non-credible content—and its propensity for impacting the quality of analytical capabilities—are all important areas in need of investigation.

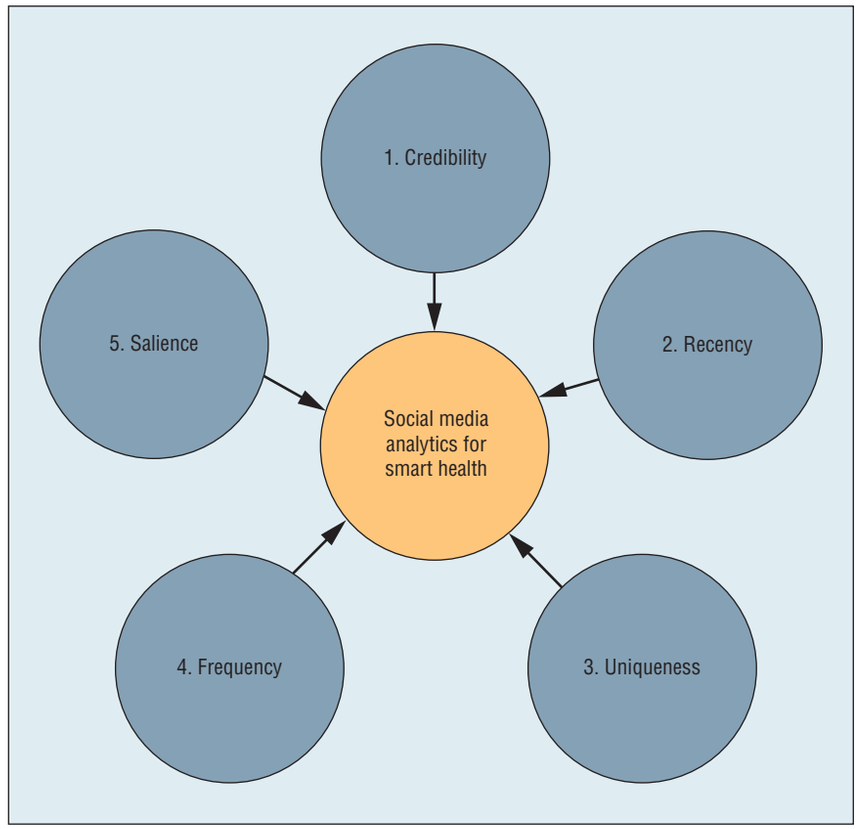

Figure 1. The credibility, recency, uniqueness, frequency, and salience (CRUFS) framework for assessing social media analytics for smart health.

Recency

When considering time series data, an important question that arises is whether indexes developed based on a particular channel constitute lead or lag indicators, both with respect to events and relative to other channels. For instance, social media was long-believed to be a lag indicator for financial and political events. However, more recently, it has been found to be an effective lead indicator.[4]

In our preliminary work, we examined the relation between social media and more than 50 drug events posted on the Food and Drug Administration (FDA) website in 2012. Three types of events were examined: product recalls, ongoing reviews, and drug safety communications. An example of a

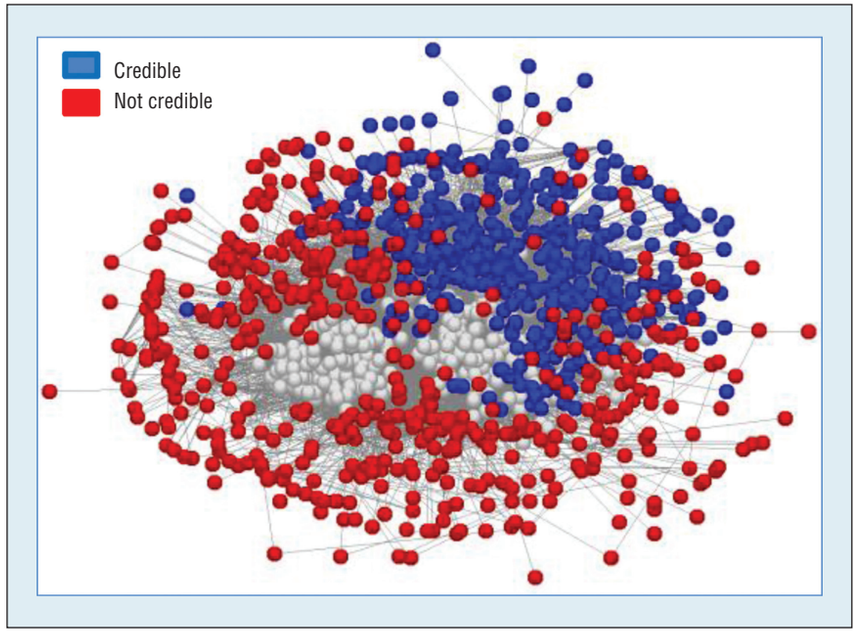

Figure 2. Credibility of select health websites.

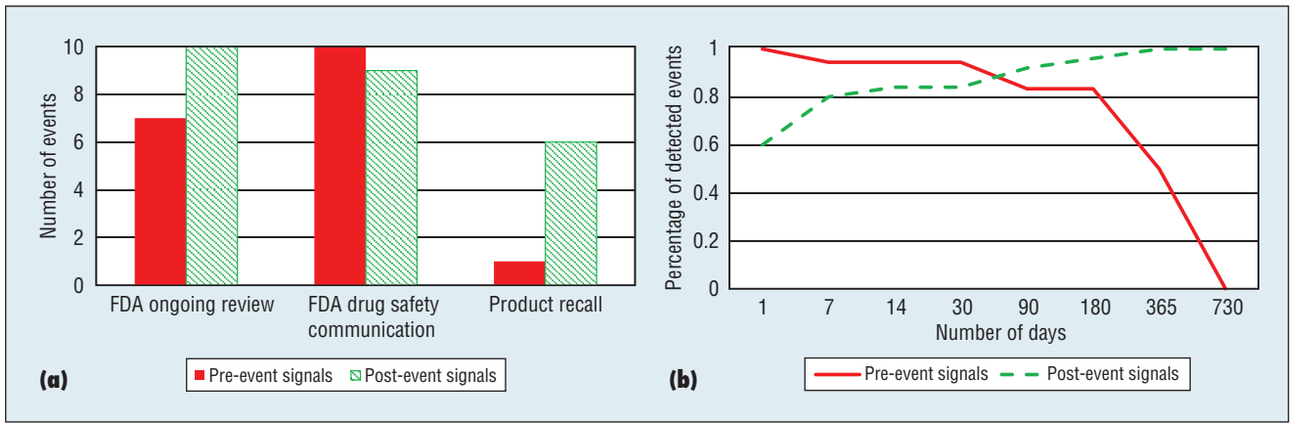

Figure 3. Examining the relation between social media and more than 50 drug events posted on the Food and Drug Administration (FDA) website in 2012. (a) Number of pre-/post-event signals detected. (b) Percentage of pre-/post-detected events for various prediction lead/lag times in days.

product recall was when Children's Tylenol was recalled due to confusion over the new dosage guidelines. Ongoing reviews are announcements of an investigation that is underway. Drug safety communications are conclusions of investigations, such as the addition of a new black-box warning for a particular drug. Figure 3 shows the results. Figure 3a shows the number of events detected prior and posterior to the FDA announcement for each category using a basic tweet mention model, where daily mention time series were constructed for each drug and spikes with a z-score greater than 3 were considered alerts. Only manually verified alerts were counted as correct event signals.

Roughly 35 percent of events had a pre-event (lead) signal, whereas 50 percent had a post-event (lag) signal. Based on Figure 3a, we can see that for certain types of events, there was a greater likelihood of pre- and post-event signals (that is, drug-safety communications), while others only had a post-event signal (such as product recalls).

Interestingly, Figure 3b shows the percentage of detected events for a given horizon in days. For pre-event signals, it shows that 83 percent of pre-event signals occurred at least 180 days prior to the first event announcement. Similarly, 80 percent of post-event signals happened within 7 days of the event. The results suggest that social media can be a strong lead or lag indicator for certain medical events, depending on the specific context. Research examining the usefulness of various channels as lead/lag indicators, across demographic segments, and for various types of events, is warranted.

Uniqueness

Channel uniqueness has important implications for data integration and fusion. The intuition behind integration in the context of analytics is epitomized by the saying "the sum is greater than its parts." Integrating data from different channels is intended to improve analytical capabilities by amplifying weak patterns across channels and combining non-overlapping "strong" patterns. However, the line between complement and duplicate can sometimes become blurred.

Consider the example presented in Figure 4. The chart shows a Twitter adverse drug-reaction-mention signal for the drug Yasmin for the years 2009–2012. The first five major signal spikes were manually annotated by examining the underlying tweets responsible for the spike. The first two events are discussions involving participants contemplating legal action due to Yasmin's side effects. The third spike is tweets/retweets of an investigative report conducted by a local news station in Oklahoma City. Similarly, the fourth and fifth are tweets/retweets triggered by an article appearing in the *British Medical Journal* and a Health Canada news release, respectively. While the first two spikes appear to be precipitated by information unique to Twitter, the remaining three are prompted by events that are also present in some other source (such as news articles, health portals, and research abstracts). Research investigating the impact of channel uniqueness and cross-channel information-diffusion patterns in the context of real-time smart health-related analytics is needed.

Frequency and Salience

Channel volumes vary considerably. On one hand, there are 1 billion new tweets every three days, 5 billion daily searches, and the number of daily forum postings and blog entries also exceeds a couple of million. Conversely, only thousands of news articles, research abstracts, and adverse event reports appear daily. Furthermore, these channels also vary with respect to the presence of salient health-related information.

Generally speaking, frequency and salience are inversely proportional. High-volume channels such as Twitter

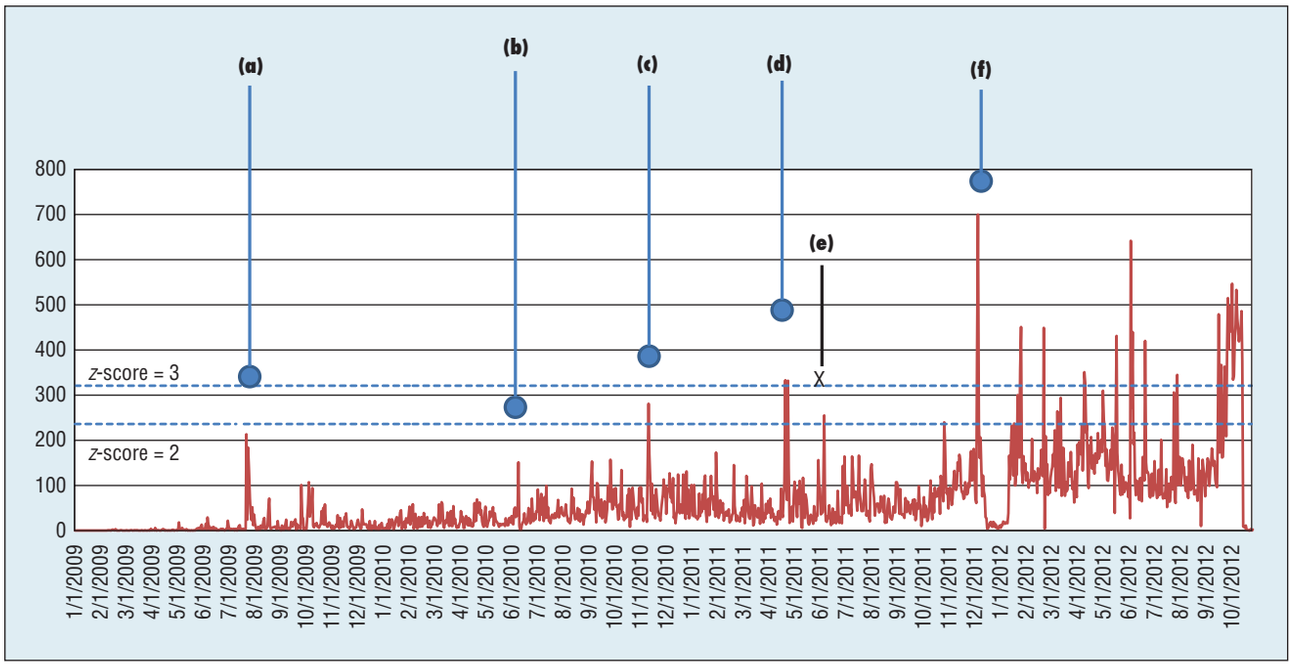

Figure 4. Charting the adverse drug-reaction-mention signal for the drug Yasmin on Twitter. (a) Discussion of lawsuits pertaining to Yasmin. (b) A Canadian woman sues Bayer. (c) An Oklahoma City investigative news local report. (d) A *British Medical Journal* article on the subject is published. (e) FDA review begins. (f) The Health Canada news release.

have average instance lengths of 20 words, thereby limiting the potential salience.[5] However, blogs and certain Web forum postings can be lengthier, with greater potential for deep content and psychometric analysis, including sentiment, affect, knowledge, and expectations. Thus, research evaluating the dynamics and implications of channel frequency and salience as they relate to real-time health analytics is necessary.

The Selected Articles and How They Relate to CRUFS

The articles appearing in this "Trends & Controversies" section constitute an excellent assortment pertaining to several key topics associated with social media analytics for smart health. They also underscore the importance of various facets of CRUFS, to varying degrees. As social media becomes a more established and mature avenue for smart-health research and practice in the future, with questions about quality, privacy, return on investment, integration, and feasibility coming front and center, we believe that CRUFS will play an even larger role in shaping the discourse and long-term viability around social media analytics for smart health.

The first article, by Mark Dredze and Michael Paul, describes the possibilities for NLP-based methods to support rich social media analytics for smart health, including modeling of health topics on Twitter, influenza surveillance, analysis of patient perceptions of physicians in Web forums, and examination of key topics in a recreational drug use discussion forum. Their work also touches upon a couple of elements of CRUFS, namely the recency and salience of social media channels such as Twitter and specialty health forums. Most notably, they point out that certain social media channels like Twitter are better-suited for general-purpose health-related analysis, while forums are better suited for more nuanced tasks, because they offer greater specificity regarding more focused topics.

Next, Fatemeh "Mariam" Zahedi and her colleagues examine the efficacy of using 3D virtual worlds with patient avatars and vital statistics (monitored using sensors) as a mechanism for delivering healthcare. Survey analytics are used to examine the effectiveness of various aspects of the system, including trust in physicians/sensors and satisfaction with the physician/group. Using actual patients with conditions such as irritable bowel syndrome and gastroesophageal reflux disease, coupled with an actual physician's avatar, they find that patients' trust and satisfaction with the virtual world delivery method is quite high. From a CRUFS perspective, future work will examine the salience of the virtual interaction cues, and the suitability of such a delivery model from not only a short-term patient's satisfaction perspective, but with respect to long-term medical diagnoses and health-related outcomes as well.

Marco D. Huesch discusses the tensions between privacy and quality of insights in the context of health-related analytics. He also touches upon the effect of health-related propaganda and misinformation, illustrated through two examples pertaining to antivaccination

campaigns. With respect to CRUFS, the work delves deep into issues surrounding credibility. For instance, he posits that health opinions lacking substantive underlying factual evidence are analogous to shouting "Fire!" in a crowded theater, blurring the lines between free speech and deception, with dire potential ramifications for social media analytics.

Finally, along with our colleagues, we examined the usefulness of combining tweets and search query volume data for early detection of adverse drug reactions. The authors examined approximately 50 events from an FDA database, and found that their signal fusion method was able to detect two-thirds of the events at least 15 months earlier (with some detected 2 to 3 years beforehand). From a CRUFS perspective, their work examines the recency, uniqueness, and frequency of drug mentions across two important channels. The work illustrates how correlated and complementary information with varying recency, frequency, and salience levels can be exploited. The authors also note the need for further work pertaining to CRUFS dimensions in the context of social media analytics for smart health.

Acknowledgments

We would like to thank the US National Science Foundation for their support through grants IIS-1236970 and IIS-1236983 entitled "Computational Public Drug Surveillance."

References

1. M. Ott, C. Cardie, and J. Hancock, "Estimating the Prevalence of Deception in Online Review Communities," *Proc. 21st Int'l Conf. World Wide Web*, 2012, pp. 201–210.
2. A. Abbasi, F.M. Zahedi, and S. Kaza, "Detecting Fake Medical Web Sites Using Recursive Trust Labeling," *ACM Trans. Information Systems*, vol. 30, no. 4, 2012, article no. 22.
3. A. Abbasi et al., "Crawling Credible Online Medical Sentiments for Social Intelligence," *Proc. ASE/IEEE Int'l Conf. Social Computing*, 2013, pp. 254–263.
4. J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *J. Computational Science*, vol. 2, no. 1, 2011, pp. 1–8.
5. A. Hassan, A. Abbasi, and D. Zeng, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework," *Proc. ASE/IEEE Int'l Conf. Social Computing*, 2013, pp. 357–364.

Ahmed Abbasi is an associate professor and Director of the Center for Business Analytics in the McIntire School of Commerce at the University of Virginia. Contact him at abbasi@comm.virginia.edu.

Donald Adjeroh is a professor and graduate coordinator in the Lane Department of Computer Science and Electrical Engineering at West Virginia University. Contact him at don@csee.wvu.edu.

Natural Language Processing for Health and Social Media

Mark Dredze and Michael J. Paul,
Johns Hopkins University

Social media, such as Twitter, has shown great potential to analyze real-world trends and events, such as politics, product sentiment, and natural disasters. In recent years, social media has emerged in the health community, particularly in public health, as a revolutionary data source for a wide range of problems. Vast amounts of naturalistic population data can be collected through social media much faster and at a lower cost than through traditional data sources such as surveys. Additionally, social media provides novel data previously unavailable to researchers. These advantages allow for the rapid formulation and evaluation of novel hypotheses, aiding decisions about how best to spend limited traditional data collection resources.

While data from social media is plentiful, it can be difficult to utilize. Most health applications query the data as if it were available as a structured database (for the example of tracking flu infections, "How many messages about flu infection on each day?"), while the data arrive in the form of unstructured text. The underlying data attributes (such as, "Does this message indicate a flu infection?") are not directly available and can only be inferred. This is the key challenge of mining health information and trends from raw text: the structure must be inferred from unstructured data, but automated methods are only viable at scale.

A basic information retrieval approach to this problem is to query for messages containing relevant keywords, such as "flu" or "fever." A limitation of this approach is that it ignores the context of these words. For example, the word "flu" could indicate a person is sick, or it might just be an acknowledgment of news articles about an ongoing flu season. The word "sick" itself has many colloquial meanings that are prevalent on social media beyond indicating illness. Methods based on *natural language processing* (NLP), an area of computer science focused on developing algorithms to understand human language, can handle these limitations by making use of richer context. Even when simple keyword querying works well, utilization of more sophisticated NLP algorithms can lead to significant improvements.[1] Additionally, for other applications, NLP provides opportunities otherwise unavailable, such as discovering health issues in social media messages, or trends in sentiment about physicians in doctor reviews.