SociAL Sensor Analytics: Measuring Phenomenology at Scale

Courtney D Corley, Chase Dowling, Stuart J Rose, and Taylor McKenzie
Knowledge Discovery and Informatics, and Visual Analytics
Pacific Northwest National Laboratory
Richland, WA, 99332
Email: {court; Chase.Dowling; Stuart.Rose}@pnnl.gov and tkmckenzie@gmail.com

Abstract—The objective of this paper is to present a system for interrogating immense social media streams through analytical methodologies that characterize topics and events critical to tactical and strategic planning. First, we propose a conceptual framework for interpreting social media as a sensor network. Time-series models and topic clustering algorithms are used to implement this concept into a functioning analytical system. Next, we address two scientific challenges: 1) to understand, quantify, and baseline phenomenology of social media at scale, and 2) to develop analytical methodologies to detect and investigate events of interest. This paper then documents computational methods and reports experimental findings that address these challenges. Ultimately, the ability to process billions of social media posts per week over a period of years enables the identification of patterns and predictors of tactical and strategic concerns at an unprecedented rate through SociAL Sensor Analytics (SALSA).

I. INTRODUCTION

Today’s analysts need methods for efficiently identifying and monitoring significant events and measuring their effects as expressed through social media. Knowing how social media is used during and in response to anticipated and unanticipated events (such as natural disasters, disease outbreaks, speeches, elections, and crises) enables more accurate measurement of the potential effect of those or similar events and informs planning and response decisions. Herein we present the findings of research on the development of a suite of social media analytics that address these challenges from the perspective that the social media user is a sensor.

Consider the process for detecting an event with a traditional sensor network (Fig. 1): 1) something happens, 2) the sensor acquires a measurement, 3) the sensor records the measurement, and 4) the system stores the measurement. Users of social media are akin to physical sensors, creating a global network of measurement capabilities, including all the inherent problems of physics-based sensors with the added complexity of spontaneous human behavior. Consider the parallel process of event responses in a social sensor network (Fig. 2): 1) something happens, 2) person receives stimulus, 3) person communicates response, 4) system routes message, 5) people receive message, 6) people communicate response, and 7) system routes message.

Research has shown that Twitter, the dominant force in what is commonly referred to as microblogging, is an excellent test bed for studying social behavior, network structures, and language use on the web systematically and quantitatively [1]–[4]. At Pacific Northwest National Laboratory we have designed and implemented a system that exploits this paradigm of considering a social media network as a sensor network by applying metadata analytics, topic clustering, and time-series modeling to a large body of commercially available social media data.

II. SOCIAL SENSOR ANALYTICS: SALSA

The SociAL Sensor Analytics (SALSA) system ingests, integrates, and makes available an immense volume of social media data with a high velocity, variety, and unknown veracity [5]. As with physical sensors, SALSA must quantify
a baseline from the social sensor measurements. A baseline provides the expected value at a particular point in time of the volume of social media features fitting some criterion, such as containing a particular hashtag, matching a complex query, or being written in a specific language or a particular dialect. SALSA quantifies baselines through metadata analytics, topic clustering, and time-series modeling.

Once a baseline is calculated, it is used to detect aberrations in the sensor data. For this research, we define an aberration as an observed value that is significantly different than expected. The method of determining the expected value provides a means of quantifying the degree or significance of the aberration. As with baselines, the value of interest is the volume of social signals that match specific criteria at a particular point in time. SALSA implements this as a brute-force approach in which potentially millions of distinct criteria are regularly evaluated (hourly or daily) with time-series models in order to quantify and rank aberrations. This allows an analyst to characterize and quantify the types of aberrations that occur during events of interest.

Ultimately, this capability illuminates the phenomenology of social media, giving power to the analyst to inform tactical and strategic reporting and planning.

A. Social Media Data in SALSA

SALSA provides immediate access to and analytical functions on over 20 billion blogs, micro-blogs, comments, and mainstream news articles spanning 13-June 2011 through 11-March 2013. The system uses PNNL’s Institutional Computing (PIC) to store and index 140 TB of social media data in a lightning-fast distributed database; the PIC is currently configured with over 600 compute nodes, a distributed Lustre file system, and a massive tape archive. SALSA retrieves these data through research access to a social media vendor application programming interface (API). Twitter data are stored in daily indexes and constitute 8.73 TB of the 140 TB. Each index stores the tweets’ 140-character text content (UTF-8) and metadata fields: unique identifier, date-hour-min-sec time stamp in UTC, author, hashtags, all mentioned users, user to which tweet was directed (@ at the beginning of the tweet), user(s) being retweeted (RT) and language (or unknown if not automatically identifiable). Of the indexed tweets 60% are in English, and the other 40% are distributed among at least 60 languages.

B. Signals from Social Sensors

Consistent with the overarching theme of social media as a sensor network, we consider the varying time-dependent measures of frequency—such as user retweets, term and hashtag usage, user-specific posts—as the social signal. Other social signals from Twitter available within SALSA are listed in TABLE I. The social signal in Fig. 3 plots the hourly frequency of hashtag “49ers” signal for the dates 13-June 2011 to 17-March 2012. The two large spikes toward the end of the time period correspond to 14-January 2012 (when the San Francisco 49ers clinched the divisional title versus the New Orleans Saints) and 22-January 2012, respectively (when the 49ers lost the NFC Champ. to the N.Y. Giants).

In addition to the social signal, we define other terminology critical to social sensor analytics. The signal magnitude is the value of the centered moving average of an indicated time period of that signal. The social signal noise is defined as the range of counts bounded by the values of two standard deviations above and below the signal magnitude. Signal aberration (or event) is an instance when the social signal exceeds signal noise boundaries. For emphasis, the terminology relevant to social sensors are defined in TABLE II.

The social signal in Fig. 4 is the hourly frequency of the hashtag “November” between the dates of 30-November 2011 (index 4000 corresponds to 14:00:00 UTC), and 5-December 2011 (index 4200 is 22:00:00). The black line is the social signal, and the red line is the signal magnitude (centered 12-hour moving average). The range of counts bounded by the upper and lower blue lines represents the signal noise boundaries (two standard deviations over and below the moving average). Here we can see that the signal overlap (black line) exceeds the signal noise (blue envelope) in what we have defined as a signal aberration (event). By inspecting Fig. 4, one can see that the signal overlap exceeds the signal noise in what is defined as a signal aberration at index 4098–30-November 2011 at 16:00:00, corresponding to 605 occurrences of the

<table>
<thead>
<tr>
<th>TABLE I. SOCIAL SIGNALS FROM TWITTER AVAILABLE WITHIN SALSA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social Signal</strong></td>
</tr>
<tr>
<td>Author frequencies</td>
</tr>
<tr>
<td>Bigram frequencies</td>
</tr>
<tr>
<td>Hashtag (and term) frequencies</td>
</tr>
<tr>
<td>Hashtag author's volume</td>
</tr>
<tr>
<td>Language frequencies</td>
</tr>
<tr>
<td>Retweet frequencies</td>
</tr>
<tr>
<td>Maximum retweet frequency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II. DEFINITIONS OF SOCIAL SENSOR TERMINOLOGY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social signal</strong></td>
</tr>
<tr>
<td>Signal magnitude</td>
</tr>
<tr>
<td>Signal noise</td>
</tr>
<tr>
<td>Signal aberration (or event)</td>
</tr>
</tbody>
</table>

FIG. 3. The figure shown is the hourly hashtag “49ers” signal for the dates 13-June 2011 to 17-March 2012. The two large spikes toward the end of the time period correspond to 14-January 2012 (when the San Francisco 49ers clinched the divisional title versus the New Orleans Saints) and 22-January 2012, respectively (when the 49ers lost the NFC Champ. to the N.Y. Giants).
hashtag “Movember.” This corresponds with the conclusion of a month-long charity event carrying the same name. (source: http://us.movember.com)

Fig. 4. The social signal in this figure is the hourly frequency of the hashtag “Movember” between the dates of 30-November 2011 (index 4000 corresponds to 14:00:00 UTC), and 5-December 2011 (index 4200 is 22:00:00). The black line is the social signal and the red line is the signal magnitude (centered 12-hour moving average). The range of counts bounded by the upper and lower blue lines represents the signal noise boundaries (two standard deviations over and below the moving average).

C. Topic Clustering in SALSA

Of particular interest to analysts and other users of social media are the related topics discussed amongst Twitter users—such as gauging the signal magnitude of politics on Twitter at any given moment. To produce baseline signals for related topics, we developed a clustering technique that uses a distance semi-metric based on author usage of the top 1,000 hashtags.

1) Mathematical Derivation of SALSA’s Distance Semi-Metric for Clustering: We first define the notion of a distance between hashtags based on author usage.

First, let $H_i^k$ be a random variable with

$$H_i^k = \begin{cases} 1 & \text{if author } i \text{ ever uses hashtag } k \\ 0 & \text{otherwise} \end{cases}$$

Also define $v_k = \langle H_1^k, H_2^k, \ldots, H_N^k \rangle$, where $N$ is the total number of authors in the sample. Further, for convenience, define $\hat{v}_k = \langle H_1^k - \overline{H_1}, H_2^k - \overline{H_2}, \ldots, H_N^k - \overline{H_N} \rangle$, where $\overline{H}$ is the sample mean of $H_i^k$ across all authors $i$. Finally, let $\mathcal{H}$ be the set of all hashtags.

Next, define the correlation between hashtags $j$ and $k$ as $\rho_{jk}$, where $\rho_{jk}$ is the Pearson correlation coefficient between $H_i^j$ and $H_i^k$. Specifically,

$$\rho_{jk} = \frac{\sum_{i=1}^{N} (H_i^j - \overline{H^j})(H_i^k - \overline{H^k})}{\sqrt{\sum_{i=1}^{N} (H_i^j - \overline{H^j})^2 \sum_{i=1}^{N} (H_i^k - \overline{H^k})^2}},$$

where $N$ is the total number of authors in the sample and $\overline{H^x}$ is the sample mean of $H_i^x$ across all authors $i$ (or, equivalently, the mean of the entries in the vector $v_x$). Also, recall the geometric interpretation of the Pearson correlation coefficient: For hashtags $j$ and $k$,

$$\rho_{jk} = \frac{\hat{v}_j \cdot \hat{v}_k}{\|\hat{v}_j\| \|\hat{v}_k\|} = \cos \theta_{jk},$$

where $\| \cdot \|$ is the vector norm and $\theta_{jk}$ is the angle between the vectors $\hat{v}_j$ and $\hat{v}_k$.

Operating on $\theta \ mod 2\pi$, we can now define $d : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ as:

$$d(j, k) = 1 - \rho_{jk}.$$ 

**Theorem.** The function $d$ is not a distance metric on the set of hashtags $\mathcal{H}$.

**Proof:** To prove that $d$ is not a metric on $\mathcal{H}$, we will provide an example of hashtag usage vectors that violate the triangle inequality for $d$. Consider hashtags $x$, $y$, and $z$ with

$$v_x = \langle 0, 0, 0, 0, 0, 1, 0, 1, 1 \rangle,$$

$$v_y = \langle 0, 1, 0, 0, 0, 0, 0, 0, 1, 1 \rangle,$$

$$v_z = \langle 0, 1, 0, 0, 0, 0, 0, 0, 1, 1 \rangle.$$ 

We can calculate the correlations between every two hashtags to be

$$\rho_{xy} = 0.8017837$$

$$\rho_{yz} = 0.8017837$$

$$\rho_{xz} = 0.5238095.$$ 

Thus, we can see that

$$d(x, y) = d(y, z) = 0.1982163$$

$$d(x, z) = 0.4761905.$$ 

Then,

$$d(x, y) + d(y, z) = 0.3964325 < 0.4761905 = d(x, z).$$ 

Thus, we have found hashtags $x$, $y$, and $z$ such that $d(x, y) + d(y, z) < d(x, z)$. As a result, there are some hashtags for which $d$ does not observe the triangle inequality, so $d$ is not a distance metric on the set of hashtags $\mathcal{H}$. ■

**Clustering Method Development**

We can define a cluster of terms $C_k^\delta$ (where $\delta$ is the clustering distance) that contains term $k$ as a recursive set:

$$C_k^\delta : \{ h \in C_k^\delta \text{ if and only if } d(h, j) < \delta \} \forall j \in \mathcal{H}.$$ 

So, $C_k^\delta$ is the set of all terms topically related (through authors) to $k$.

Equivalently, for a term $k$, we can let $B_\delta(k)$ be a ball of radius $\delta$ around $k$. That is,

$$B_\delta(k) = \{ h \in \mathcal{H} \mid d(h, k) < \delta \}.\footnote{It should be noted that a term cluster is a collection of terms relating to a single common topic if and only if authors only tweet about a single topic. This may not be far from reality given that many people only discuss a small number of topics. Regardless, this clustering method was developed to determine which authors are influential to other authors, and the distinction of topic areas is not necessary for this analysis.}$$
For a term \( h \in \mathcal{H} \), if \( B_\delta(k) \cap B_\delta(h) \neq \emptyset \), then they are a part of the same connected component. Further, if for \( j, h \in \mathcal{H} \),
\[
B_\delta(k) \cap B_\delta(h) \neq \emptyset,
\]
\[
B_\delta(h) \cap B_\delta(j) \neq \emptyset,
\]
and
\[
B_\delta(k) \cap B_\delta(j) = \emptyset,
\]
the term \( j \) is still in the connected component of \( k \) since there is a connection through \( B_\delta(h) \). Then we can define a cluster of terms \( C^\delta_k \) as the set of all terms \( h \in \mathcal{H} \) such that \( h \) is part of the largest connected component containing \( k \).

2) Effects on Topic Clustering of Varying Distance Thresholds: A social signal was formed by the most-used 1,000 hashtags by author from the period 13-June 2011 to 17-March 2012. Then the Pearson correlation was determined between every pair of hashtags, distances between hashtags were calculated using the previously described semi-metric, and clusters were constructed using the definition above with a clustering distance of \( \delta = 0.80 \). TABLE III enumerates the hashtags from several topic clusters calculated using this parameters. At this distance, the algorithm identifies 70 non-singleton clusters.

<table>
<thead>
<tr>
<th>Subject Area of Cluster</th>
<th>Hashtags Identified in Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Media</td>
<td>Apple, Business, Facebook, Google, Marketing, SEO, SocialMedia, Twitter, business, iPad, iPhone, market-ing, mobile, seo, social, socialmedia</td>
</tr>
<tr>
<td>Socio-political Activism</td>
<td>#14Feb, #14Feb, Anonymous, BAHRAIN, BBC, Bahrain, CNN, China, EU, Egypt, Feb14, France, GCC, GOP, Gaza, Hamas, India, Iran, Iraq, Islam, Israel, Jan25, KSA, Kuwait, Libya, London, NDAA, OWS, Obama, Occupy, OccupyOakland, OccupyWallStreet, PIPA, Pakistan, Palestine, Qatar, Qurban, RonPaul, SOPA, Saudi, Syria, Tehran, UAE, UK, UN, US, USA, Yemen, aluminium, bahrain, economy, egypt, feb14, gov, humanrights, iran, islam, jan25, ksa, kuwait, occupy, occupywallstreet, ocr, ows, p2, politics, q8, saudi, sgp, syria, tehrir, teaparty, tlot, uk, usa</td>
</tr>
<tr>
<td>Justin Bieber Fan Media</td>
<td>1DFact, 1DFamily, 1DQuotes, BELIEVE, BOYFRIEND, Beliebers, Believe, BieberFact, DREAMBIG, ILOVENYMANS, Imagine, LISTEN, MISTLETOE, MUCHLOVE, Mistletoe, NEVERSAYNEVER, NEW-MUSIC, NOV1st, NSN, NeverSayNever, REAL, SOMEDAY, SWAG, Someday, Twitition, UNDER-THE-MISTLETOE, UnderTheMistletoe, askilam, beliebers, believe, bieberfact, charity, epic, grateful, imagine, makeachange, mistletoe, machlove, neversaynever, real, someday, swag, twitatsunday, underthehmisule-toe</td>
</tr>
<tr>
<td>U.S. Professional Sports</td>
<td>Eagles, Giants, Jets, MLB, NBA, NFL, Patriots, Rangers, WorldSe ries, n, postseason</td>
</tr>
<tr>
<td>Health and Fitness</td>
<td>Health, diet, fitness, food, health, weightloss</td>
</tr>
</tbody>
</table>

The clusters in TABLE III appear to correlate what we might perceive to be similarly related hashtags. We do, however, find that a significant number of hashtags are identified as singleton clusters (487 of 1000). Indicated in Fig. 5, is the number of clusters identified as the maximum distance between hashtags is relaxed from 0.0 to 1.0 in increments of 0.01. The number of clusters identified converges exponentially to one large cluster. Interestingly, at a near-maximum cluster distance of 0.99 (1.00 implies a single cluster) the following top-ranking hashtags remain in singleton clusters based on author usage: “Rift”, “Valor”, “bakugetki”, “cidade_rio”, “magistream”, and “pokeveganow”. Of these “magistream” appears to be a type of social media game. “Bakugetki” and “pokeveganow” are largely associated with the Japanese language and “cidade_rio” with the Portuguese language, according to Google Translate. “Rift” appears to be a video game and “Valor” was not readily associated with any particular topic [http://twitter.com/search]. Given that SALSA is clustering based on author usage, it seems likely that the English, Japanese, and Portuguese hashtags appearing in the top 1,000 hashtags would have very little user overlap. We suspect that the difference in message content language—implying little similarity in top 1,000 hashtag usage per author—is governing this end-point behavior. This would explain why these clusters remain singletons at cluster distance values close to 1.0.

Nevertheless, with this technique we are now able to determine the signal magnitude of related topics by summing the users’ hashtags’ signals: see the example for health-related topics. In this example, Fig. 6 plots the hourly social signal from the “health” cluster (hashtags: Health, diet, fitness, food, health, weightloss) between the dates 13-June 2011 to 17-March 2012. The health social signal has a mean of 277.4 occurrences per hour and a standard deviation of 92.77.

![Fig. 5. Variation of the Number of Clusters with Change in Clustering Distance. The x-axis labels the cluster distance threshold. The y-axis labels the cluster size.](image)

![Fig. 6. The hourly social signal from the “health” cluster (hashtags: Health, diet, fitness, food, health, weightloss) between the dates 13-June 2011 to 17-March 2012. The health social signal has a mean of 277.4 occurrences per hour and a standard deviation of 92.77.](image)

D. Social Signal Modeling to Inform Phenomenology

Yet to be developed is a theory for social media’s phenomenology composed of patterns and signatures of events and topics over a period of time. The class of time-series models that SALSA uses is the Autoregressive Integrated Moving Average (ARIMA) [6]–[11]. We first define our response
variable $Y_t$ as the social signal frequency at hour $t$. Let $I_t$ be the frequency of terms (of the impact of terms as a strictly increasing function of the number of people that would see the usage of those terms) related to the social signal used by accounts deemed to be influential (related) at hour $t$, and let $\epsilon_t$ be independent and identically distributed random variables with $\epsilon_t \sim N(0, \sigma^2)$ for all $t$. We assume that $Y_t$ can be made stationary by some degree of differencing; let $d$ be the least such degree of differencing and let $D_t$ be $Y_t$ differenced by degree $d$. Then our model states that

$$D_t = \sum_{i=1}^{p} \psi_i D_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + c + \epsilon_t,$$

where $\psi_i$, $\theta_i$, and $c$ are model parameters. Equivalently, we say that $Y_t$ is an ARIMA model of order $(p, d, q)$. The model orders are chosen to maximize the log likelihood of the goodness of fit to the time series. Further, we also consider the signal magnitude (centered moving average) and the corresponding signal noise (moving standard deviation) over time periods of 12, 24, 48, 72, and 168 hours and model these resultant time series independently [12].

1) Topic Clusters and ARIMA Model Performance: With respect to interest in modeling signals relevant to the user message space, we sought to evaluate model performance for all hashtags at all clustering distances. We evaluate model performance (fit) by determining the correlation between the measured signal values and those values minus the ARIMA model’s value’s residuals. For each cluster, the model fit was iteratively calculated between the range 0.0 and 1.0 at 0.01 increments. The model orders and correlations were then used to select the “best” cluster signals or all cluster signals at that distance. At a maximum cluster distance of 0.0, all hashtag clusters are guaranteed to be singletons, thereby observing individual hashtag signal model performance as well as observing hashtag cluster model performance at more relaxed distances.

We calculate both the average model performance as well as the standard deviation of model performances for each of the cluster distances. The calculations show there is non-significant change in average model performance as clustering distances are relaxed. In Fig. 7 one can see the minimal change in average model performance until maximum cluster distance exceeds .90—where the number of identified clusters begins to quickly converge to one. We find a small local maximum in average model performance, .577, at a distance of .80. After this optimization calculation, SALSA fixed the cluster distance at .80, and for the 1,000 most-used hashtags, 665 clusters (including singletons) were found.

2) Signal Aberration Detection: In an attempt to reduce the effects of the significant variation of the signal from hour to hour, we then calculate the signal moving average and its moving standard deviation. This allows us to capture the average signal magnitude for a specified time period but also associate the degree of variance for that time period—what we are referring to as signal noise. Considering this moving average independent of the associated moving standard deviation allows us to model the signal magnitude and the signal noise separately.

At a clustering distance of .80, we calculated each cluster’s signal magnitude (12-hr and 24-hr moving average) and noise boundaries (moving standard deviation). We employed the ARIMA model selection algorithm and method of measuring model fit and find that the model of the 12-hr moving standard deviation—the 12-hr signal noise—exhibits an average 92.78% model fit over all 665 hashtag clusters with a variance of 0.2%. The model of the 12-hr moving average—the 12-hr signal magnitude—exhibits an average of 0.76% with a variance of 0.3%. For the 24-hr signal noise and signal magnitude, we find average model fits of 96.95% and .04% with variances of .03% and .61%, respectively.

III. DISCUSSION AND CONCLUSION

The foundation of Pacific Northwest National Laboratory’s SociAL Sensor Analytics (social signals, signal aberrations and events) is driven in part by the need to characterize the phenomenology of social media and specifically Twitter, as indicated in previous work such as Yang and Leskovec [13]. SALSA’s aim is not to predict the next signal aberration, but rather identify the conditions under which temporal phenomenologies manifest and to associate phenomenologies with event types. It is through this lens that SALSA aims to predict event types before or as they manifest—whether they are responses to a spontaneous unanticipated event (e.g., natural disaster), the result of an information campaign (e.g., public health messaging), or reactions to other viral topics. This mode of analysis could be used to consider signal aberration in the long run—topic popularity, disinformation, user censorship, etc.

SALSA clusters topics in a manner that uses the content and user as the units of analysis. SALSA implements topic clustering through using the dot product similarity metric between authors and their hashtag usage, over the course of a specified time period. Within each cluster, the sum of the individual hashtags determines the cluster volume (likewise...
for frequencies). There are several pitfalls: hashtags are merely a proxy for relevant message topics. Hashtags are terse (and not always literal) representations of the message topic at hand and are subject to frequent changes in usage over time. Moreover, this method of clustering is static and provides clusters of hashtag usage by author over the specified time period. This method’s accuracy of classification beyond the dates specified is severely limited by the measurement of the 1,000 most-used hashtags and their use by individual authors of that time period. This method of clustering, however, serves as a successful proof-of-principle in eliciting social signals from topic categories. The clustering results provide actionable information in spite of the usage of a naïve brute-force algorithm.

The time-series representations of social signals are modeled to study social signals and inform phenomenology. The features may be studied individually or in an ensemble to calibrate the social signal that may be attenuated by exogenous factors. There are myriad potential signals that could be measured from the data stream, e.g. posts per hour directed from user to user (@s and mentions), highest volume users per hour, or most influential users per day. Determining the markers for these social signal aberrations such that we can anticipate emergent event types—not predicting that a meteor is about to crash, but predicting whether or not a signal aberration is the result of a random event or an information campaign.

ACKNOWLEDGMENTS

This work was funded by the Pacific Northwest National Laboratory’s National Security Directorate through the Laboratory Directed Research and Development Program. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.

REFERENCES


