Analyzing Social Media Content for Security Informatics

Richard Colbaugh
Sandia National Laboratories
Albuquerque, NM USA
colbaugh@comcast.net

Kristin Glass
Sandia National Laboratories
Albuquerque, NM USA
klglass@sandia.gov

***Abstract*—Inferring public opinion regarding an issue or event by analyzing social media content is of great interest to security analysts but is also technically challenging to accomplish. This paper presents a new method for estimating sentiment and/or emotion expressed in social media which addresses the challenges associated with Web-based analysis. We formulate the problem as one of text classification, model the data as a bipartite graph of documents and words, and construct the sentiment/emotion classifier through a combination of semi-supervised learning and graph transduction. Interestingly, the proposed approach *requires no labeled training documents* and is able to provides accurate text classification using only a small lexicon of words of known sentiment/emotion. The classification algorithm is shown to outperform state of the art methods on a benchmark task involving sentiment analysis of online consumer product reviews. We illustrate the utility of the approach for security informatics through two case studies, one examining the possibility that online sentiment about suicide bombing predicts bombing event frequency, and one investigating public sentiment about vaccination and its implications for population health and security.**

***Keywords*—**-text analysis, social media, sentiment, emotion, machine learning, security informatics.

I. Introduction

Web-users regularly offer their views and opinions concerning security-relevant topics, such as emerging conflicts or potential epidemics, and there is considerable interest in leveraging this information to support various objectives (e.g., threat warning, disease surveillance [1-4]). Moreover, there is evidence for the feasibility of this idea. For instance, there has been substantial recent work in non-security domains demonstrating the utility of analyzing Web data to support both forecasting [5-9] and monitoring/surveillance [10-15] missions. While social media analytics targeting security applications is less common, recent research shows these data can be helpful for early warning [16-18] and situational awareness [19,20,3].

Public sentiment and emotion regarding issues and events, and the way it is distributed and evolving, is of particular interest in many settings, such as when assessing the threat posed by a contentious situation or likelihood that a new vaccine will be adopted. The capability to infer public sentiment/emotion from Web content is appealing because it could permit such assessments to be performed rapidly and inexpensively. Unfortunately this inference problem presents many challenges, and we mention three that have been especially daunting.

First, the *volume* of user-generated Web content is enormous (e.g., Facebook has over 1B active users and Twitter users produce more than 350M posts per day [21]); thus exploring and analyzing these data demands very efficient algorithms. Next, most social media content is expressed using *informal and imprecise language,* suggesting that only algorithms which can learn and adapt have the potential to be useful. Indeed, previous research has shown that machine learning (ML) is well-suited to such tasks [22] and may provide a good framework for conducting sentiment analysis (see [23] for a general review and [24] for a security-oriented discussion). ML is not a panacea, of course, and this leads to the third challenge: ML achieves efficiency and flexibility by learning from examples, and this implies that *labeled examples* must be available for algorithm training. Obtaining labeled data in security applications is typically expensive and time-consuming [24].

In this paper we consider the problem of inferring sentiment or emotion expressed in social media content, for example a blog post. It is of interest to estimate sentiment/emotion toward a specific topic along each of four “axes”: 1.) sentiment (positive, negative), 2.) affective valence (pleasant, unpleasant), 3.) arousal (excited, calm), and 4.) dominance (dominating, dominated) [25]. Furthermore, motivated by the nature of security informatics, we seek the capability to infer sentiment/emotion for new classes of content (e.g., written in an unfamiliar language or about a novel emerging topic) without the need to collect and label new training documents.

This paper presents a new method for estimating sentiment or emotion which addresses the challenges associated with social media analysis. We formulate the problem as one of text classification, model the data as a bipartite graph of documents and words, and construct the sentiment/emotion classifier using a combination of semi-supervised learning and graph transduction. The classifier can be implemented *with no labeled training documents,* and enables accurate text classification using only a modest lexicon of words of known sentiment/emotion polarity. The proposed algorithm is shown to outperform gold standard methods on a benchmark sentiment analysis task involving online consumer product reviews. Additionally, we illustrate the utility of the methodology for security informatics through two case studies, one that relates online public opinion concerning suicide bombing to frequency of bombing events, and one which indicates that sentiment about vaccination can affect immunization decisions of individuals and therefore adversely impact the health and security of a population.

II. Preliminaries

The goal of this paper is to develop an accurate, flexible, scalable, and easy-to-implement procedure for estimating the sentiment and/or emotion of social media content. We approach this task as one of document classification. Motivated by the requirements of security informatics applications, we assume the sentiment/emotion classifier must be learned using only a small lexicon of words of known polarity, and that no labeled documents are available for training. This approach eliminates the need to collect and label example documents and to retrain the classifier each time a new domain is encountered, making it better suited to security-oriented analysis than standard content analysis methods.

We wish to infer the polarity of a collection of documents toward a specific topic along one or more of the following four axes: sentiment, valence (or "happiness"), arousal, and dominance [25]. Each document of interest is represented as a "bag of words" feature vector $\mathbf{x}\in\Re^{|V|}$, where the entries of $\mathbf{x}$ are the frequencies with which the words in a vocabulary set V appear in the document (perhaps normalized in some way [23]). One way to estimate the orientation of document $\mathbf{x}$ with respect to the sentiment/emotion axis currently under examination is to learn a vector $\mathbf{c}\in\Re^{|V|}$ such that the classifier orient = sign($\mathbf{c}^{T}\mathbf{x}$) returns +1 if $\mathbf{x}$ is 'positive', 'happy', 'excited', or 'dominating' and −1 if instead $\mathbf{x}$ is 'negative', 'unhappy', 'calm', or 'dominated'.

It is assumed that a lexicon of words of known polarity is available for use in learning our document classifier. That is, we suppose we have acquired lexicons of positive words $V^{+}\subseteq V$ and negative words $V^{-}\subseteq V$, where 'positive' and 'negative' are defined in the context of the task at hand; thus 'positive' refers to positive sentiment, happy, excited, or dominating words, and 'negative' to their opposites. Note that high-quality lexicons characterizing the four axes of interest are publicly available. For example, the "IBM lexicon" is a collection of 2968 words which have been assigned {positive, negative} sentiment labels by the IBM India Research Labs (for other text analysis applications [26]). As for emotion content, the Affective Norms for English Words (ANEW) lexicon consists of 1034 words that have been assigned numerical scores with respect to valence (happiness), arousal, and dominance [25]. In what follows, for convenience of presentation the analytic task of interest will be referred to as *sentiment* analysis, but it should be clear that the methodology we derive is also directly applicable to analysis of valence, arousal, and dominance.

Observe that a simple way to construct a document classifier of the form orient = sign($\mathbf{c}^{T}\mathbf{x}$) is to specify $\mathbf{c}$ by setting $c_i = +1$ if word $i\in V^{+}$, $c_i = -1$ if $i\in V^{-}$, and $c_i = 0$ if i is not in either lexicon. Such a classifier simply sums the positive and negative polarity words in the document and assigns document orientation accordingly. While this "lexicon-only" scheme can yield acceptable performance in narrowly-focused domains, it is usually labor-intensive to build lexicons that are sufficiently complete to yield useful document classification. Alternatively, standard ML-based methods attempt to compute classifier vector $\mathbf{c}$ from examples of positive and negative sentiment. Learning often is based upon a set of labeled documents $\{d_i, \mathbf{x}_i\}^{nl}_{i=1}$, where $d_i\in\{+1, -1\}$ is the sentiment label for document i. While this approach can be effective [23], existing learning methods require that large training sets of labeled documents be collected [23].

Sentiment analysis of social media content for security applications is often characterized by the existence of only modest levels of prior knowledge regarding the domain of interest, reflected in the availability of a small lexicon of sentiment-laden words, and by the need to rapidly learn and adapt to new situations. Consequently, standard sentiment analysis methods are typically ill-suited for security informatics. To address this challenge, we propose in [24] an algorithm which leverages an additional source of data which is abundant online: *unlabeled* documents and words. Specifically, the algorithm assumes the existence of a corpus of n documents, of which $n_l << n$ are labeled, and a modest lexicon of sentiment-laden words, and combines these labeled data with the information present in the $n - n_l$ unlabeled documents via semi-supervised learning. In the present work we extend this result, deriving a sentiment classifier which provides good performance in settings where *no* labeled documents are available.

The development of sentiment classifiers which effectively combine information contained in labeled and unlabeled data is facilitated by modeling the problem data as a bipartite graph G_b of documents and words (see Figure 1). It is easy to see that the adjacency matrix for the graph G_b is given by

$$A = \begin{bmatrix} 0 & X \\ X^{T} & 0 \end{bmatrix} \qquad (1)$$

where the matrix $X\in\Re^{n\times|V|}$ is constructed by stacking the document vectors as rows, and each '0' is a matrix of zeros. This bipartite graph model permits labeled and unlabeled data to be integrated by exploiting the relationships between documents and words encoded in G_b via the *clustering hypothesis:* it is assumed that positive (labeled or unlabeled) documents tend to be connected to positive (labeled/unlabeled) words, and analogously for negative documents and words. Because it is assumed that we have only a small number of labeled words and no labeled documents, the process of clustering on graph G_b will be important to the success of the approach.

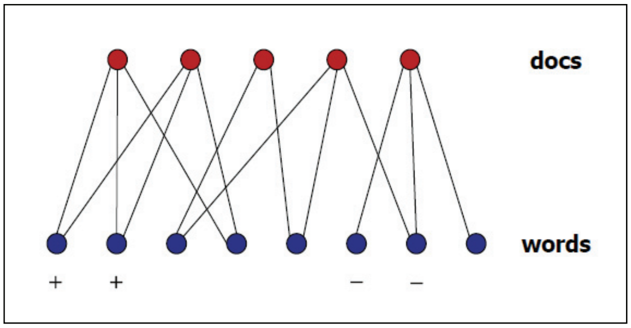

Figure 1. Cartoon of bipartite graph model G_b, in which documents (red vertices) are connected to the words they contain (blue vertices). The polarities of the words included in the lexicon $V_l=V^{+}\cup V^{-}$ are indicated with + / − symbols.

III. Sentiment/Emotion Analysis

A. Sentiment/Emotion Estimation Algorithm

The problem of interest may be stated as follows: given a corpus of n documents of unknown sentiment or emotion orientation and a modest lexicon $V_l = V^+ \cup V^- \subseteq V$ of words of labeled (sentiment/emotion) polarity, estimate the orientation of all of the documents. We adopt a ML classification approach, and are interested in both *document classification,* which estimates the sentiment polarity of a particular document, and *corpus classification,* where the goal is to infer the sentiment of a collection of documents. The latter task seeks assessments of the form: "In this set of blog posts, X% of bloggers feel positively about topic Y". Estimates of this sort are frequently of great interest to security analysts. Interestingly, it is often possible to achieve higher accuracy when assessing a corpus than when analyzing an individual document, because some of the false-positive and false-negative classification errors cancel each other in corpus-level assessments. (For instance, this partial cancellation of errors occurs in the corpus-level classification test described in Section IIIB).

The proposed approach to sentiment classification consists of three main steps, which can summarized as follows:

1. Construct an "intermediate" sentiment classifier C: document → class using the semi-supervised ML algorithm derived in [24].
2. Employ classifier C to estimate the sentiment orientation of all the documents. Assign "preliminary" labels to those documents about which C is "confident".
3. Obtain the final document polarity estimates by performing graph transduction [27] on the partially-labeled bipartite graph data model built in Step 2.

Of course, this procedure must be specified more carefully in order to be implementable. Step 1 involves learning a sentiment classifier C via the methodology developed in [24]. The initial problem data consists of a corpus of n documents and a small lexicon of words with known sentiment labels; this label information is encoded as vector $\mathbf{w} \in \Re^{|Vl|}$, where $V_l = V^+ \cup V^-$ as before and each component of $\mathbf{w}$ is set to +1 if the corresponding word belongs to V^+ and −1 if the word is in V^-. Let $\mathbf{d}_{est} \in \Re^n$ denote the vector of estimated sentiment orientations for the documents in the corpus, and define an "augmented" classifier $\mathbf{c}_{aug} = [\mathbf{d}_{est}^T \quad \mathbf{c}^T]^T \in \Re^{n+|V|}$ which estimates the polarity of both documents and words. In [24] the quantity $\mathbf{c}_{aug}$, and therefore $\mathbf{c}$, is learned by solving an optimization problem, and $\mathbf{c}$ is used to estimate the sentiment of any document $\mathbf{x}$ via the linear classifier orient = sign($\mathbf{c}^T\mathbf{x}$).

More specifically, the algorithm derived in [24] learns classifier $\mathbf{c}_{aug}$ in such a way that it possesses two properties: 1.) if a word is contained in the sentiment lexicon V_l then the corresponding entry of $\mathbf{c}$ is close to this ±1 polarity, and 2.) if there is an edge X_{ij} of G_b that connects a document $\mathbf{x}$ and a word $v \in V$ and X_{ij} has significant weight, then the estimated orientations of $\mathbf{x}$ and v are similar. These objectives can be encoded in a minimization problem as follows:

$$\min_{c_{aug}} \; c_{aug}^T L_n^k c_{aug} + \beta \sum_{i=1}^{|V_l|} (c_i - w_i)^2 \qquad (2)$$

where L = D − A is the graph Laplacian matrix for G_b (thus D is diagonal with $D_{ii} = \Sigma_j A_{ij}$), $L_n = D^{-1/2} L D^{-1/2}$ is the normalized graph Laplacian, k is a positive integer, and β is a nonnegative constant. The $\mathbf{c}_{aug}$ which minimizes objective function (2) can be obtained by solving a set of sparse linear equations [24], where the equations are sparse because data matrix X is sparse. Thus large problems can be solved efficiently (e.g. via the Conjugate Gradient method) and the sentiment classifier is scalable to large datasets.

In Step 2, classifier C is first employed to estimate the sentiment polarity of all documents in the corpus, and then preliminary polarity labels are assigned to those documents about which C is "confident". More precisely, preliminary labels d_i, $i \in \{1, \ldots, n_l\}$, are assigned to the n_l documents with large magnitude sentiment polarity estimates: for nonnegative parameter d_{thresh}, if $d_{est,i} > d_{thresh}$ ($d_{est,i} < -d_{thresh}$) then $d_i = +1$ ($d_i = -1$), and the document indices are reordered (if necessary) so that the first n_l are the ones assigned these preliminary labels. Graph G_b is then updated to G'_b by incorporating these n_l document labels.

Finally, Step 3 of the process computes the final document sentiment orientation estimates $\mathbf{d}_{est,f} \in \Re^n$ via graph transduction [27] on the updated bipartite graph G'_b (i.e., the graph with n_l preliminary document labels). Our graph transduction scheme propagates the n_l document labels and $|V_l|$ word labels to all the documents and words in graph G'_b by solving the following minimization problem:

$$\min_{c_{aug}} \; c_{aug}^T L_n^k c_{aug} + \beta_1 \sum_{i=1}^{n_l} (d_{est,f,i} - d_i)^2 + \beta_2 \sum_{i=1}^{|V_l|} (c_i - w_i)^2 \quad (3)$$

where now $\mathbf{c}_{aug} = [\mathbf{d}_{est,f}^T \quad \mathbf{c}^T]^T$ and β_1, β_2 are positive constants. Note that the preliminary document labels $\mathbf{d} = [d_1 \ldots d_{nl}]^T$ and word labels $\mathbf{w}$ are not strictly enforced; rather, deviations of $\mathbf{d}_{est,f}$ from $\mathbf{d}$ and $\mathbf{c}$ from $\mathbf{w}$ are penalized in the objective function (3). This approach to graph transduction appears to enjoys performance advantages over other formulations (see the empirical tests reported in Section IIIB).

We summarize this discussion by sketching an algorithm for learning the proposed sentiment/emotion estimation (SEE) classifier.

Algorithm SEE:

1. Construct intermediate classifier C: orient = sign($\mathbf{c}^T\mathbf{x}$) by solving optimization problem (2) for $\mathbf{c}_{aug} = [\mathbf{d}_{est}^T \quad \mathbf{c}^T]^T$.
2. Employ classifier C to estimate the sentiment orientation of all documents. Assign preliminary labels $\mathbf{d} \in \Re^{nl}$ to the n_l documents with large magnitude polarity estimates (i.e., those documents for which $|d_{est,i}| > d_{thresh}$).
3. Compute the final sentiment orientation estimates $\mathbf{d}_{est,f}$ for all of the documents through graph transduction by solving optimization problem (3) for $\mathbf{c}_{aug} = [\mathbf{d}_{est,f}^T \quad \mathbf{c}^T]^T$.

The performance of Algorithm SEE is now evaluated empirically through sentiment analysis of a collection of online consumer product reviews.

B. Empirical Evaluation

We now assess the performance of Algorithm SEE for the task of estimating the sentiment of consumer product reviews. The dataset used in the investigation is a collection of 2000 online reviews for electronics products, 1000 positive and 1000 negative, archived at the website [28]. The Lemur Toolkit [29] was employed to construct the data matrix X from these reviews. A lexicon of 150 domain-independent sentiment-laden words, 75 positive and 75 negative, was assembled manually from the IBM lexicon [26] and used to form the vector **w** of word labels. Observe that, given the existence of publicly-available lexicons such as those described in [25,26], building an appropriate **w** is usually convenient.

This study compares the sentiment classification accuracy of Algorithm SEE with that of two other algorithms: a *lexicon-only* (LO) strategy and the *structural correspondence learning* (SCL) sentiment classifier proposed in [30]. Algorithm SEE is implemented with the following parameter values: Step 1: β = 0.5, k = 5; Step 2: d_{thresh} is set so that n_l = 200 (i.e., 200 documents are assigned preliminary polarity labels); Step 3: β_1 = 0.1, β_2 = 0.5, k = 3. Recall that the LO strategy estimates sentiment orientation of document **x** according to the formula orient = sign($\mathbf{c}^T\mathbf{x}$), where **c** has nonzero entries corresponding to the 150 word lexicon described above. To ensure an informative comparison, the SCL sentiment estimator is applied exactly as described in [30], where it was tuned for the same electronics reviews dataset as is used here. Note that the SCL algorithm requires labeled documents for training; in the present study the training/testing procedure consisted of training on a random sample of 1600 labeled product reviews and then testing on the remaining 400 "held-out" reviews. Because the SCL method is shown in [30] to outperform other classifiers and has access to substantial information unavailable to Algorithm SEE, the SCL algorithm can be viewed as a gold standard for this task.

Sample results from this study are displayed in Figure 2. Sentiment classification accuracy is estimated via standard ten-fold cross-validation [22], and the accuracy values reported in the plot represent the average of ten trials. It is seen that the document classification accuracies for Algorithm SEE, the LO strategy, and the SCL classifier are 85.0%, 70.8%, and 84.4%, respectively. Thus Algorithm SEE, which uses the same labeled information as the LO method, performs slightly better than the gold standard SCL classifier, which is trained on 1600 labeled reviews. We also applied Algorithm SEE to the corpus classification problem. Because the algorithm is relatively unbiased, its performance at the corpus level is quite strong: the proportion of positive and negative reviews in the set is estimated with 95.3% accuracy. It is expected that the ability of Algorithm SEE to provide good sentiment estimation accuracy using only a small lexicon of sentiment-laden words and no labeled documents will be of considerable value in security applications. To explore this possibility, we now consider two security-oriented case studies.

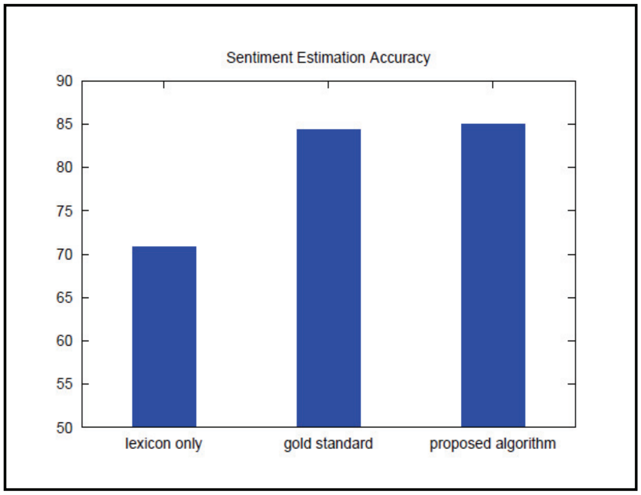

Figure 2. Results for empirical test on consumer product reviews. The bar chart shows that sentiment classification accuracy of the lexicon-only method, gold standard classifier, and Algorithm SEE are 70.8%, 84.4%, and 85.0%, respectively. Notice that the accuracy of Algorithm SEE is slightly higher than that of the gold standard method, despite the fact that the latter requires much more labeled information for training.

IV. Case Study One: Suicide Bombing

Recent research indicates that certain characteristics of online discourse about contentious situations can be predictive of real-world behavior. For instance, [16] shows that specific patterns in the diffusion of blog discussions regarding protests and cyber attacks provide practically-useful early warning for such events. Motivated by this research, the present case study considers the following question: is regional online sentiment concerning the use of violence predictive of violent activity? Observe that such a relationship is plausible. Extremist elements in a society often depend upon that society for financial and other support, as a pool from which to recruit new participants, and as a safe haven for planning and operations [31], so that these groups may respond, implicitly or explicitly, to local public opinion. However, there has been little empirical work done to examine this possibility.

As a concrete instantiation of this question, we explored the relationship between regional online sentiment about Palestinian suicide bombing attacks against Israel and actual bombing events. We collected two forms of data for the period 1997-2006: 1.) Arabic-language online content (mainly blog posts and editorials) discussing Palestinian suicide bombing against Israel, and 2.) descriptions of suicide bombing attacks carried out against Israel by four Palestinian groups: Fatah, Hamas, the Popular Front for Liberation of Palestine, and the Palestinian Islamic Jihad. (The period 1997-2006 is the longest for which both online content and real-world attack data could be collected.)

The sentiment regarding suicide bombing expressed in the online content was estimated using Algorithm SEE. This classifier was implemented exactly as described in Section III, with the exception that lexicon V_1 was obtained through interviews with two domain experts [31] and then manually translated into Arabic (the elicitation and translation procedure took approximately four hours). Data on suicide bombings was assembled from multiple sources (see [31] for a description of all of the raw data). Time series for bombing sentiment and bombing events are plotted in Figure 3. The blue curve depicts the fraction of posts that express positive sentiment concerning suicide bombings against Israel, and the red curve shows the (normalized) frequency of such attacks by the groups Fatah, Hamas, PFLP, and PIJ.

Visual examination of the plots presented in Figure 3 indicates that the dynamics of sentiment and bombings are related. Analysis of the data confirms this impression: online sentiment about suicide bombing and frequency of bombing attacks are correlated, with sentiment leading event frequency by twelve months (Pearson correlation coefficient $r = 0.8$, $p < 0.0001$). This finding suggests the intriguing possibility that online discourse regarding some types of violence may be predictive of violent activity.

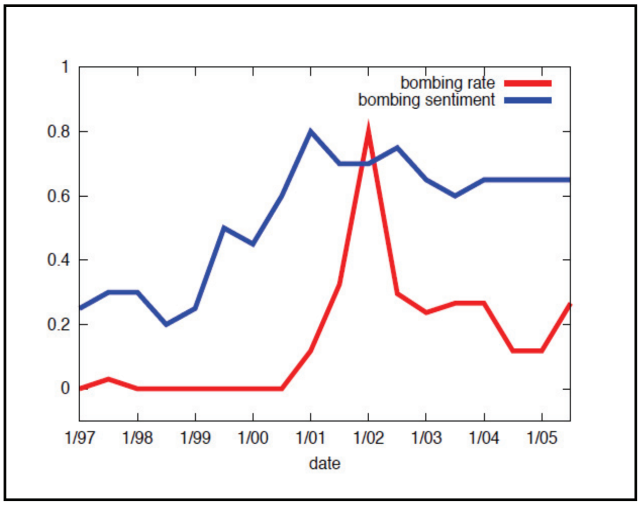

Figure 3. Results for suicide bombing case study. The plots show that regional online sentiment supporting Palestinian suicide bombing attacks against Israel (blue curve) is correlated with bombing frequency (red curve, normalized), with changes in sentiment leading changes in event frequency by twelve months (Pearson coefficient $r = 0.8$, $p < 0.0001$).

V. Case Study Two: H1N1 Vaccination

Public sentiment regarding health issues may affect behavior, which in certain circumstances would lead to increased health risks and decreased security in a population. As an illustrative and important example, sentiment concerning vaccination has been shown to impact individual vaccination decisions, giving rise to the possibility of large epidemics [32,4,14]. Disease propagation has been studied for many decades, and much is known about the emergence and control of epidemics [e.g. 33]. However, it is only recently that researchers have begun to analyze the role of public sentiment in epidemics and to propose ways to leverage social media to monitor this sentiment and mitigate health risk [32,14,18].

This case study provides a preliminary investigation of such issues by examining the Influenza A H1N1 pandemic of 2009 [32]. We collected ~500M Twitter posts made by ~17M users during the seven month period from 1 June to 31 December 2009; this period was selected for study because it contains the “Fall wave” of the H1N1 pandemic. These data were preprocessed to extract two main quantities: 1.) the content of each post, modeled as a bag-of-words vector $\mathbf{x} \in \Re^{|V|}$ (for vocabulary V, see Section II), and 2.) a Twitter graph $G_T = (V_T, E_T)$, where vertices $v \in V_T$ are users and edges $(v_1,v_2) \in E_T$ connect users who have communicated via @-messages. More precisely, we place an edge $(A,B) \in E_T$ from user A to user B if A has included ‘@B’ in at least three posts during the seven month study period (other threshold choices yield similar graphs).

With the preprocessed data in hand, we first identified those posts that are related to H1N1 vaccination through a keyword search, employing the keywords given in [32]. The sentiment of these posts was then estimated using Algorithm SEE. The algorithm was implemented as described in Section III, with the exception that lexicon V_1 was augmented through the addition of a few common ‘emoticon’ symbols. Estimating the sentiment of the full corpus of relevant posts reveals that sentiment toward H1N1 vaccination is positive on average, with slightly more than 60% of posts being classified as positive (Figure 4, green bar).

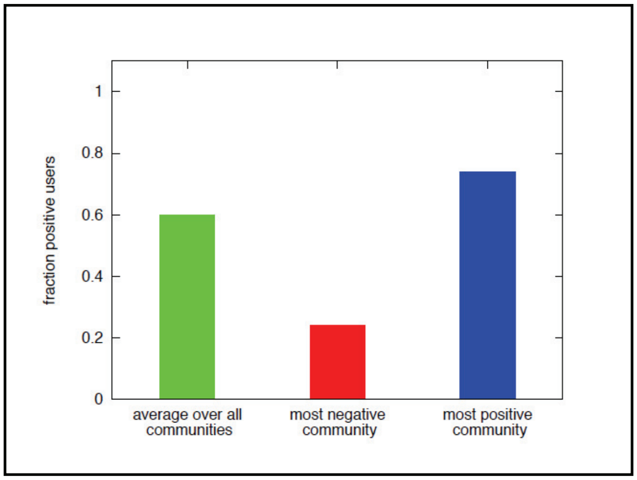

Figure 4. Sentiment analysis results for H1N1 vaccination case study. The bar chart shows that sentiment regarding vaccination, while positive on average (green bar), is distributed heterogeneously over the Twitter graph (red and blue bars), and that some communities express quite negative sentiment (red bar).

Next we examined the possibility that sentiment concerning H1N1 vaccination might be distributed heterogeneously over the Twitter graph, with individuals clustering into groups according to their opinions about vaccination efficacy and risks. While there are numerous ways to divide a graph into groups, perhaps the most natural is to partition it into *graph communities,* that is, densely connected groupings of individuals which have relatively few links between them. Adopting this definition for groups, we applied the graph community partitioning process described in [16] to the *reduced* Twitter graph formed by users who made at least one post about H1N1 vaccination; this computational procedure identified 23 graph communities. Then, for each community, all H1N1 vaccination-relevant posts generated by users belonging to that community were assembled, and Algorithm SEE was used to estimate the sentiment of those posts.

As can be seen from Figure 4, this analysis shows that sentiment is indeed distributed heterogeneously across the Twitter graph's communities. For instance, in the most positive community ~73% of posts are positive, while in the most negative community only ~24% of posts are positive. Observe that this result could have important implications for public health. For example, if a heterogeneous distribution of sentiment about vaccines leads to a clustered distribution of vaccination status in the population, communities with low vaccination rates may not be protected by herd immunity even if the overall vaccination rate is high.

We explored this issue in a preliminary way by performing a simulation study of the impact of vaccination sentiment clustering on epidemic risk. In the simulations, sentiment clustering is assumed to: 1.) be identical to that displayed in Figure 4, and 2.) give rise to a comparable distribution of vaccination status. Disease dynamics are simulated within the multi-scale hybrid dynamical system framework presented in [5]. The epidemic propagates with susceptible-exposed-infectious-recovered dynamics, and the social network over which the disease spreads is modeled as possessing the same topology as Twitter graph G_T.

The effects of vaccination sentiment clustering on epidemic risk was evaluated by running two sets of 100 simulations. In the first set, 60% of individuals are chosen uniformly at random from the population and vaccinated. Thus, in these runs, the vaccination rate matches the proportion of the population with positive vaccination sentiment, but the distribution of vaccinations is assumed to be homogeneous. In contrast, in the second set of simulations, individuals are vaccinated according to the vaccination sentiment distribution found in the Twitter graph. Thus, in this set of runs, 60% of individuals are again vaccinated, but the vaccination probability varies heterogeneously over the communities (see Figure 4). All simulations are initiated by infecting a small number of randomly chosen individuals.

Sample results are shown in Figure 5. Each bar represents the average number of infected individuals over 100 simulation runs, with the red and blue bars corresponding to homogeneous and heterogeneous vaccination distributions, respectively (note that infection rates are normalized to permit convenient comparison). It can be seen that populations with clustered distributions of vaccination status, arising from similar distributions of vaccination sentiment, experience substantially larger epidemics than those in which vaccination status is distributed homogeneously (more than 2.5 time larger on average).

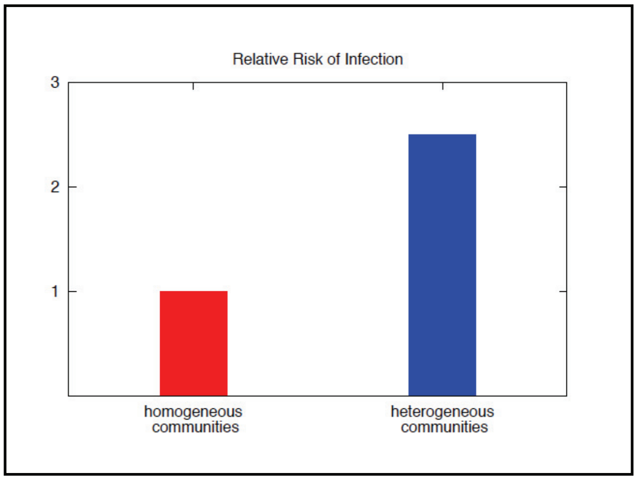

Figure 5. Simulation results for H1N1 vaccination case study. The bar chart shows that populations with clustered distributions of vaccination status (blue bar) experience substantially larger epidemics than those for which vaccination status is distributed homogeneously (red bar).

VI. Concluding Remarks

This paper presents a new method for estimating the sentiment or emotion of social media content which is accurate and also "agile", in that the computational algorithm is implementable *with no labeled training documents*. The proposed approach is shown to outperform gold standard techniques on a benchmark task involving sentiment analysis of online consumer product reviews. Additionally, the utility of the methodology for security informatics is illustrated through two case studies, one examining the predictive power of online sentiment about suicide bombing and one investigating the way sentiment about vaccination is distributed over the population and the implications of this distribution for public health.

Future work will include: 1.) comprehensive exploration of the potential of the approach for multi-lingual sentiment analysis, 2.) examination of the impact of document length on text analysis (e.g. because security-relevant social media posts, such as Tweets on Twitter, are often quite terse), and 3.) application of the approach to additional health-related security informatics problems.

Acknowledgements

This research was supported in part by the Laboratory Directed Research and Development Program at Sandia National Laboratories. We thank Travis Bauer, Nancy Brodsky, Robert Glass, Philip Kegelmeyer, and David Zage of Sandia for many stimulating discussions, and Prem Melville of IBM Research for sharing with us the "IBM lexicon".

REFERENCES

[1] Chen, H. et al., "Uncovering the Dark Web: Case study of jihad on the Web", *J. American Society for Information Science and Technology*, Vol. 59, pp. 1347-1359, 2008.

[2] Friedland, J. and K. Rogerson, "Literature review: How political and social movements form on the Internet and how they change over time", Institute for Homeland Security Solutions Report, November 2009.

[3] Glass, K. and R. Colbaugh, "Web analytics for security informatics", *Proc. 2011 EISIC,* Athens, Greece, September 2011.

[4] Salathe, M. et al., "Digital epidemiology", *PLoS Computational Biology,* Vol. 8 (7), 2012.

[5] Colbaugh, R. and K. Glass, "Predictive analysis for social processes I: Multi-scale hybrid system modeling, and II: Predictability and warning analysis", *Proc. 2009 IEEE MSC*, Saint Petersburg, Russia, July 2009.

[6] Asur, S. and B. Huberman, "Predicting the future with social media", *Proc. WI-IAT '10*, Toronto, Canada, September 2010.

[7] Berger, J., A. Sorensen, and S. Rasmussen, "Positive effects of negative publicity: When negative reviews increase sales", *Marketing Science,* Vol. 29, pp. 815-827, 2010.

[8] Bollen, J., H. Mao, and X. Zeng, "Twitter mood predicts the stock market", *J. Computational Science*, Vol. 2, pp. 1-8, 2011.

[9] Amodea, G., R. Blanco, and U. Brefeld, "Hybrid models for future event prediction", *Proc. CIKM '11*, Glasgow, Scotland, UK, October 2011.

[10] Lawrence, R. et al., "Social media analytics", *OR/MS Today,* pp. 26-30, February 2010.

[11] Maniu, S., B. Cautis, and T. Abdessalem, "Building a signed network from interactions in Wikipedia", *Proc. DBsocial '11*, Athens, Greece, June 2011.

[12] Ayers, J., K. Ribisi, and J. Brownstein, "Tracking the rise in popularity of electronic nicotine delivery systems using search query surveillance", *American J. Preventative Medicine*, Vol. 41, pp. 1-6, 2011.

[13] Colbaugh, R. and K. Glass, "Leveraging sociological models for prediction I: Inferring adversarial relationships, and II: Early warning for complex contagions", *Proc. 2012 IEEE ISI*, Washington, DC USA, June 2012.

[14] Bhattacharya, S. et al., "Belief surveillance with Twitter", *Proc. WebSci 2012*, Evanston, IL USA, June 2012.

[15] Kosinski, M., D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior", *Proc. National Academy of Sciences USA*, Early Edition, pp. 1-4, March 2013.

[16] Colbaugh, R. and K. Glass, "Early warning analysis for social diffusion events", *Security Informatics,* Vol. 1 (18), 2012.

[17] Zammit-Mangion, A. et al., "Point process modelling of the Afghan War Diary", *Proc. National Academy of Sciences USA*, Vol. 109, pp. 12414-12419, 2012.

[18] Sadilek, A., H. Kautz, and V. Silenzio, "Predicting disease transmission from geo-tagged micro-blog data", *Proc. 26th AAAI Conference on AI*, Toronto, Canada, July 2012.

[19] Abbasi, A., H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", *ACM Transactions on Information Systems*, Vol. 26, pp. 1-34, 2008.

[20] Lampos, V., T. De Bie, and N. Cristianini, "Flu detector – Tracking epidemics on Twitter, *ECML PKDD 2010*, Springer LNAI 6323, 2010.

[21] http://www.wikipedia.org , accessed March 2013.

[22] Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Second Edition, Springer, New York, 2009.

[23] Pang, B. and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2, pp. 1-135, 2008.

[24] Colbaugh, R. and K. Glass, "Agile sentiment analysis of social media content for security informatics applications", *Proc. 2011 EISIC,* Athens, Greece, September 2011.

[25] Bradley, M. and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings", Technical Report C1, University of Florida, 1999.

[26] Ramakrishnan, G. et al., "Question answering via Bayesian inference on lexical relations", *Proc. 41st Annual Meeting ACL*, Sapporo, Japan, July 2003.

[27] Belkin, M., P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples", *J. Machine Learning Research*, Vol. 7, pp. 2399-2434, 2006.

[28] http://www.cs.jhu.edu/~mdredze/, accessed December 2010.

[29] http://www.lemurproject.org/, accessed December 2010.

[30] Blitzer, J., M. Dredze, and F. Perieia, "Biographies, bollywood, boom-boxes, and blenders: Domain adaptation for sentiment classification", *Proc. 45th Annual Meeting ACL*, Prague, June 2007.

[31] Colbaugh, R. et al., "TIARA", Sandia Report SAND2009-0325, Sandia National Laboratories, June 2010.

[32] Signorini, A., A. Segre, and P. Polgreen, "The use of Twittter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic", *PLoS One,* Vol. 6 (5), 2011.

[33] Anderson, R. and R. May, *Infectious Diseases of Humans: Dynamics and Control,* Oxford University Press, UK, 1992.