Estimating Sentiment Orientation in Social Media for Intelligence Monitoring and Analysis

Richard Colbaugh
Sandia National Laboratories and New Mexico Tech
colbaugh@nmt.edu

Kristin Glass
New Mexico Tech
kglass@icasa.nmt.edu

***Abstract*—This paper presents a computational approach to inferring the sentiment orientation of "social media" content (e.g., blog posts) which focuses on the challenges associated with Web-based analysis. The proposed methodology formulates the task as one of text classification, models the data as a bipartite graph of documents and words, and uses this framework to develop a semi-supervised sentiment classifier that is well-suited for social media domains. In particular, the proposed algorithm is capable of combining prior knowledge regarding the sentiment orientation of a few documents and words with information present in unlabeled data, which is abundant online. We demonstrate the utility of the approach by showing it outperforms several standard methods for the task of inferring the sentiment of online movie reviews, and illustrate its potential for security informatics through a case study involving the estimation of Indonesian public sentiment regarding the July 2009 Jakarta hotel bombings.**

***Keywords*—**-sentiment analysis, social media, security informatics.

I. Introduction

There is increasing recognition that the Web represents a valuable source of national security-relevant intelligence and that computational analysis offers a promising way of dealing with the problem of collecting and analyzing data at Web scale [e.g., 1-3]. As a consequence, tools and algorithms have been developed which support various security informatics objectives [1]. To cite a specific example, we have shown that blog network dynamics can be exploited to provide reliable early warning for a class of extremist-related, real world protest events [4].

Making sense of online content at Web scale is both important and technically challenging [5]. For instance, discussions on social media sites such as blogs and forums often reflect the sentiments and opinions of individuals and groups about security-relevant topics, and thus can represent valuable intelligence data. However, these views are typically expressed as informal communications and are buried in vast volumes of irrelevant discourse, so that efficiently and accurately extracting them is usually quite difficult [e.g., 5-9]. While powerful analysis techniques have been derived for traditional forms of content, less has been done to develop strategies that are well-suited to the particular characteristics of content generated in social media.

Consider the important task of deciding whether a given post expresses positive or negative opinion toward a topic of interest. The informal, multilingual nature of social media content poses a challenge for language-based sentiment analysis. While statistical learning-based methods often provide good performance in unstructured settings like this [e.g., 5-9], obtaining the required labeled instances of data, such as a lexicon of sentiment-laden words for a given domain or a collection of "exemplar" blog posts of known polarity, is labor-intensive and time-consuming for Web applications.

This paper presents a new approach for inferring sentiment orientation of social media content which addresses the challenges associated with Web-based analysis. We formulate the task as one of text classification, and assume availability of a few documents of known polarity and small lexicon of words about which we have sentiment knowledge. Additionally, we suppose the existence of a set of unlabeled documents, and note that such data are readily obtainable in online applications. The proposed approach models the data as a bipartite graph of documents and words, and uses this framework to develop a semi-supervised sentiment classifier that is capable of combining the prior knowledge reflected in the labeled documents and words with information present in the unlabeled data. We demonstrate the utility of the algorithm by showing that it outperforms several other methods for the task of inferring the sentiment of online movie reviews, and illustrate the potential of the approach for security informatics by using it to analyze blog content and estimate Indonesian public sentiment toward the alleged planner of the July 2009 Jakarta hotel bombings.

II. Problem Formulation

We approach the task of estimating sentiment orientations of a collection of documents as a text classification problem. Each document of interest is represented as a "bag of words" feature vector $x \in \Re^{|V|}$, where the entries of x are the frequencies with which the words in the vocabulary set V appear in the document. We wish to learn a vector $c \in \Re^{|V|}$ such that the classifier orient = $\text{sign}(c^T x)$ accurately estimates the orientation of document x, returning +1 (−1) for documents expressing positive (negative) sentiment about the topic of interest.

Knowledge-based classifiers leverage prior domain information to construct the vector c. One way to obtain such a classifier is to assemble lexicons of positive words $V^+ \subseteq V$ and negative words $V^- \subseteq V$, and then to set $c_i = +1$ if word $i \in V^+$, $c_i = -1$ if $i \in V^-$, and $c_i = 0$ if i is not in either lexicon; this classifier simply sums the positive and negative sentiment words in the document and assigns document orientation accordingly. While this scheme can provide acceptable performance in certain settings, it is unable to improve its performance or adapt to new domains, and it is usually labor-intensive to construct lexicons which are sufficiently complete to enable useful sentiment classification performance to be achieved.

Alternatively, learning-based methods attempt to generate the classifier vector c from examples of positive and negative sentiment. To obtain a learning-based classifier, one can begin by assembling a set of n_l *labeled* documents $\{(x_i, d_i)\}$, where $d_i \in \{+1, -1\}$ is the sentiment label for document i. The vector c then can be learned through "training" with the set $\{(x_i, d_i)\}$, for instance by solving the following set of equations for c:

$$[X^T X + \gamma I_{|V|}]\, c = X^T d, \tag{1}$$

where matrix $X \in \Re^{nl \times |V|}$ has document vectors for rows, $d \in \Re^{nl}$ is the vector of document labels, $I_{|V|}$ denotes the $|V| \times |V|$ identity matrix, and $\gamma \geq 0$ is a constant; this corresponds to regularized least squares (RLS) learning. Many other strategies can be used to compute c, including Naïve Bayes (NB) statistical inference [5]. Learning-based classifiers have the potential to improve their performance and to adapt to new situations, but realizing these capabilities requires that training sets of labeled documents be obtained and this is usually an expensive undertaking.

National security-related sentiment analysis applications are often characterized by the existence of only modest levels of prior knowledge, reflected in the availability of a few labeled documents and small lexicon of sentiment-laden words, and by the need to rapidly learn and adapt to new domains. This is particularly true in the case of Web-based security informatics, where the wide range of domains and languages of interest and importance of timeliness combine to make analytic responsiveness a central issue. To address these challenges, the sentiment analysis method proposed in this paper leverages an important feature of Web-based analysis: it is easy to collect large sets of *unlabeled* documents and words on the Web.

We assume available a collection of n documents, of which $n_l << n$ are labeled, and a modest lexicon of sentiment-laden words. The lexicon is encoded as a vector $w \in \Re^{|Vl|}$, where $V_l = V^+ \cup V^-$ is the sentiment lexicon and the entries of w are set to +1 or −1 according to their polarity. We model the problem data as a bipartite graph G_b of documents and words. It is easy to see that the adjacency matrix A for graph G_b is given by

$$A = \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \tag{2}$$

where each '0' is a matrix of zeros. Integration of labeled and unlabeled data is accomplished by exploiting the relationships between documents and words encoded in the bipartite graph. The basic idea is to assume that, in the bipartite graph G_b, positive/negative documents will tend to be connected to (contain) positive/negative words, and positive/negative words will tend to be connected to (contained in) positive/negative documents.

III. Sentiment Estimation Method

Assume given a corpus of n documents, of which $n_l < n$ are labeled, and a modest lexicon V_l of sentiment-laden words; suppose that this label information is encoded as vectors $d \in \Re^{nl}$ and $w \in \Re^{|Vl|}$, respectively. Let $d_{est} \in \Re^n$ be the vector of estimated sentiment orientations for the documents in the corpus, and define the "augmented" classifier $c_{aug} = [d_{est}^T \;\; c^T]^T \in \Re^{n+|V|}$ which estimates the polarity of both documents and words. Note that the quantity c_{aug} is introduced for notational convenience in the subsequent development and is not directly employed for classification. More specifically, our approach is to learn c_{aug}, and therefore c, by solving an optimization problem involving the labeled and unlabeled training data, and then to estimate the sentiment of any new document of interest with the simple linear classifier orient = $\mathrm{sign}(c^T x)$. We assume that documents and words are indexed so the first n_l elements of d_{est} and $|V_l|$ elements of c correspond to the labeled data.

We wish to learn an augmented classifier c_{aug} with the following three properties: 1.) if a document is labeled, then the corresponding entry of d_{est} should be close to this ±1 label; 2.) if a word is in the sentiment lexicon, then the corresponding entry of c should be close to this ±1 sentiment polarity; and 3.) if there is an edge X_{ij} of G_b that connects a document x and a word $v \in V$ and X_{ij} possesses significant weight, then the estimated polarities of x and v should be similar. These objectives are encoded in the following minimization problem:

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

where $L = D - A$ is the graph Laplacian matrix for G_b, with D the diagonal degree matrix for A (i.e., $D_{ii} = \Sigma_j A_{ij}$), and β_1, β_2 are nonnegative constants. Minimizing (3) enforces the three properties we seek for c_{aug}, with the second and third terms penalizing "errors" in the first two properties. To see that the first term enforces the third property, note that this expression is a sum of components of the form $X_{ij}\,(d_{est,i} - c_j)^2$. The constants β_1, β_2 can be used to balance the relative importance of the three properties.

The c_{aug} which minimizes objective function (3) can be obtained by solving the following set of linear equations:

$$\begin{bmatrix} L_{11} + \beta_1 I_{nl} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} + \beta_2 I_{|V_l|} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix} c_{aug} = \begin{bmatrix} \beta_1 d \\ 0 \\ \beta_2 w \\ 0 \end{bmatrix} \tag{4}$$

where the L_{ij} are matrix blocks of L of appropriate dimension. This system is very sparse because the data matrix X is sparse, and therefore large-scale problems can be solved efficiently. Note that in situations where the set of available labeled documents and words is *very* limited, we have found that sentiment classifier performance can be improved by replacing L in (4) with the normalized Laplacian $L_n = D^{-1/2} L D^{-1/2}$, or a power of this matrix L_n^k (where k is a positive integer).

We summarize by sketching the proposed algorithm:

Algorithm SO:

1. Construct (4), possibly by replacing the graph Laplacian L with L_n^k.
2. Solve (4) for $c_{aug} = [\, d_{est}^T \;\; c^T \,]^T$ (for instance using the Conjugate Gradient method).
3. Estimate the sentiment orientation of any document x of interest as: orient = $\mathrm{sign}(c^T x)$.

IV. Case Study 1: Movie Reviews

The first case study examines the performance of Algorithm SO for the problem of estimating sentiment of online movie reviews (known to be a difficult task). The data used in this study is a publicly available set of 2000 movie reviews, 1000 positive and 1000 negative [10]. A lexicon of ~1400 domain-independent sentiment-laden words was obtained from [11] and employed to build the lexicon vector w. The study compares the sentiment estimation accuracy of Algorithm SO with that of three other schemes: 1.) lexicon-only, in which lexicon vector w is used as the classifier as summarized in Section II, 2.) a standard NB classifier obtained from [12], and 3.) a well-tuned RLS classifier (1). Algorithm SO is implemented with the following parameter values: $\beta_1 = 0.1$, $\beta_2 = 0.5$, and $k = 10$.

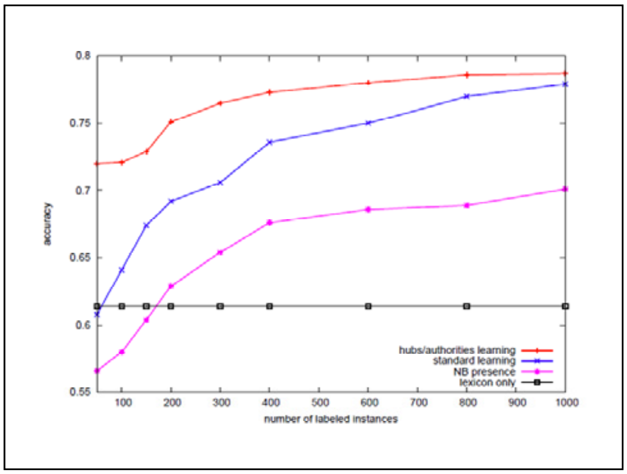

Figure 1. Results for movie reviews case study. The plot shows how sentiment estimation accuracy (vertical axis) varies with number of available labeled movie reviews (horizontal axis) for four different classifiers: lexicon only (black), NB (magenta), RLS (blue), and Algorithm SO (red).

Sample results of this study are depicted in Figure 1. In each trial, the movie reviews are randomly partitioned into 1000 training and 1000 test documents, and a randomly selected subset of training documents of size n_l is "labeled" (i.e., labels for these reviews are made available to the learning algorithms). Observe that Algorithm SO outperforms the other three methods., particularly when the number of labeled training documents is small.

V. Case Study 2: Blogs

On 26 July 2009 a document claiming responsibility for the 19 July Jakarta hotel bombings and allegedly written by Noordin Top was posted on the blog [13]. At the time, senior U.S. intelligence and security officials expressed interest in understanding sentiment in the region regarding the bombings and the alleged claim of responsibility by a well-known extremist. To enable a preliminary assessment along these lines, we collected two sets of social media data related to the Top post: 1.) the ~3000 comments made to the post during the two week period immediately following its publication, and 2.) several hundred posts made to other Indonesian language blogs in which the Top post was discussed. We manually labeled the sentiment of a small subset of these documents, and also translated into Indonesian the generic sentiment lexicon employed in Case Study 1 for use in this study. We implemented Algorithm SO to estimate the sentiment expressed in the corpus of comments made to the blog [13] and in the set of related discussions posted at other blogs. This analysis revealed that the comments made to the Top post are almost universally negative, condemning both the bombings and the justification for the bombings given in the Top post. Manual examination of a subset of the comments confirms the results provided by Algorithm SO. Analysis of relevant posts made to other blogs also express largely negative sentiments about the Top post, although they are not as consistently negative as the comments made directly on the blog site [13].

Acknowledgements

This research was supported by the U.S. Department of Defense and the Laboratory Directed Research and Development program at Sandia National Laboratories.

References

[1] Chen, H., C. Yang, M. Chau, and S. Li (Editors), *Intelligence and Security Informatics*, Lecture Notes in Computer Science, Springer, Berlin, 2009.

[2] US Committee on Homeland Security and Government Affairs, *Violent Extremism, the Internet, and the Homegrown Terrorism Threat*, 2008.

[3] Bergin, A., S. Osman, C. Ungerer, and N. Yasin, "Countering Internet Radicalization in Southeast Asia", ASPI Special Report, March 2009.

[4] Colbaugh, R. and K. Glass, "Predictive Analysis for Social Processes I: Multi-Scale Hybrid System Modeling, and II: Predictability and Warning Analysis", *Proceedings 2009 IEEE Multi-Conference on Systems and Control*, Saint Petersburg, Russia, July 2009.

[5] Pang, B. and L. Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, Vol. 2 , pp. 1-135, 2008.

[6] Das, S. and M. Chen, "Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards", *Proceedings APFA*, 2001.

[7] Dhillon, I., "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning", *Proceedings ACM International Conference on Knowledge Discovery and Data Mining*, San Francisco, August 2001.

[8] Kim, S. and E. Hovy, "Determining the Sentiment of Opinions", *Proceedings International Conference on Computational Linguistics*, 2004.

[9] Sindhwani, V. and P. Melville, "Document-Word Co-Regularization for Semi-Supervised Sentiment Analysis", *Proceedings 2008 IEEE International Conference on Data Mining*, Pisa, Italy, December 2008.

[10] http://www.cs.cornell.edu/People/pabo/movie-review-data/.

[11] Ramakrishnan, G., A. Jadhav, A. Joshi, S. Chakrabarti, and P. Bhattacharyya, "Question Answering via Bayesian Inference on Lexical Relations", *Proceedings ACL*, 2003.

[12] http://www.borgelt.net/bayes.html.

[13] www.mediaislam-bushro.blogspot.com.