SAND2011-6399C

SAND2011-6399 C

Agile Sentiment Analysis for Social Media Content

Rich Colbaugh*† **Kristin Glass**†

*Sandia National Laboratories
†New Mexico Institute of Mining and Technology

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

September 2011

Introduction

Objective

Develop an effective, scalable computational methodology for estimating sentiment orientation in "social media" (e.g., blogs, networking sites).

Motivation

Discussions on social media sites often reflect sentiments and opinions of individuals and groups about security-relevant topics.

Challenges

- Sentiment is typically *expressed informally* and buried in vast volumes of irrelevant discourse.
- *Labeling* exemplars of positive/ negative documents and words is *expensive and time-consuming*.

Introduction

Outline

- Problem formulation:
  - text classification;
  - bipartite graph data model.
- Semi-supervised approach:
  - algorithm;
  - sample results.
- New approach:
  - algorithm;
  - sample results.
- National security applications.

Problem Formulation

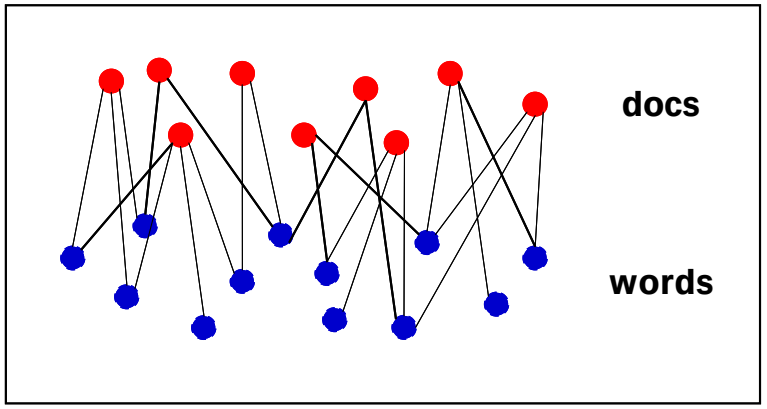
Sentiment analysis as text classification

- Setup: construct vector $c \in \Re^{|V|}$ so that classifier sign($c^T x$) accurately estimates sentiment of "bag-of-words" document vectors $x \in \Re^{|V|}$ (V is vocabulary).
- Standard methods:
  - knowledge-based (e.g., use lexicon of sentiment-laden words to construct c) – unable to improve performance or adapt to new situations;
  - learning-based (e.g., use set of labeled documents to learn c) – able to improve and adapt but expensive to label documents.
- Proposed approach: use learning with limited labeled examples, and supplement labeled data with *unlabeled* examples (which are abundant online).

Problem Formulation

Bipartite graph data model

- Assume given:
  - corpus of n documents, of which $n_l << n$ are labeled ($d \in \Re^{nl}$) – note that $n_l = 0$ is possible);
  - modest lexicon V_l of sentiment-laden words ($w \in \Re^{|Vl|}$).

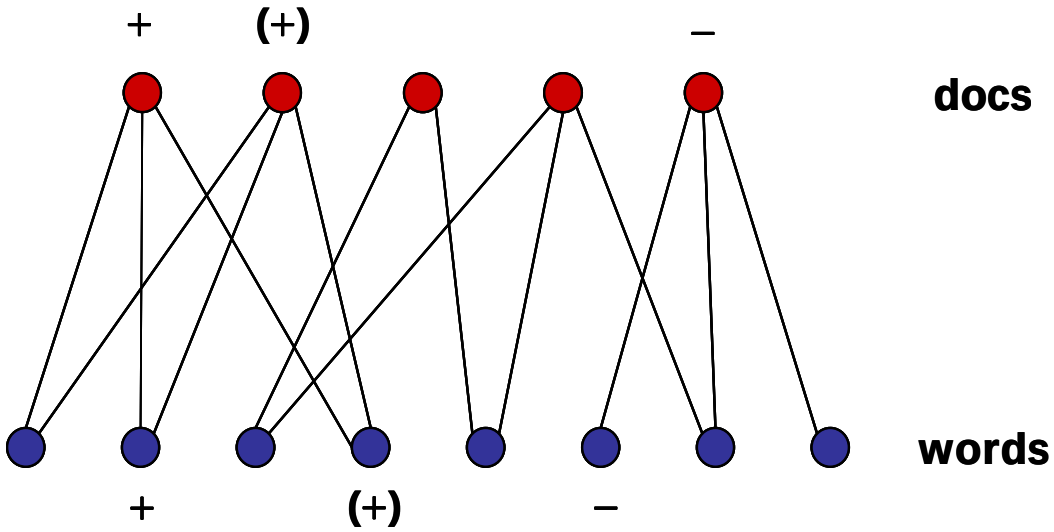


- Analytic approach: leverage information in *unlabeled* documents by:
  - modeling data as a bipartite graph G_b of documents and words;
  - assuming that, in G_b, positive/negative documents (words) will tend to be connected to positive/negative words (documents).

Semi-Supervised Learning

Background

- Basic idea:



- Mathematics:

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

where $c_{aug} = [d_{est}^{T} \;\; c^{T}]^{T}$ are doc/word sentiment estimates, $L = D - A$ is graph Laplacian for G_b, and $c_{aug}^{T} L \, c_{aug}$ is sum of $X_{ij}(d_{est,i} - c_j)^2$ terms.

Semi-Supervised Learning

Algorithm SS

1. Construct the following set of linear equations:

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

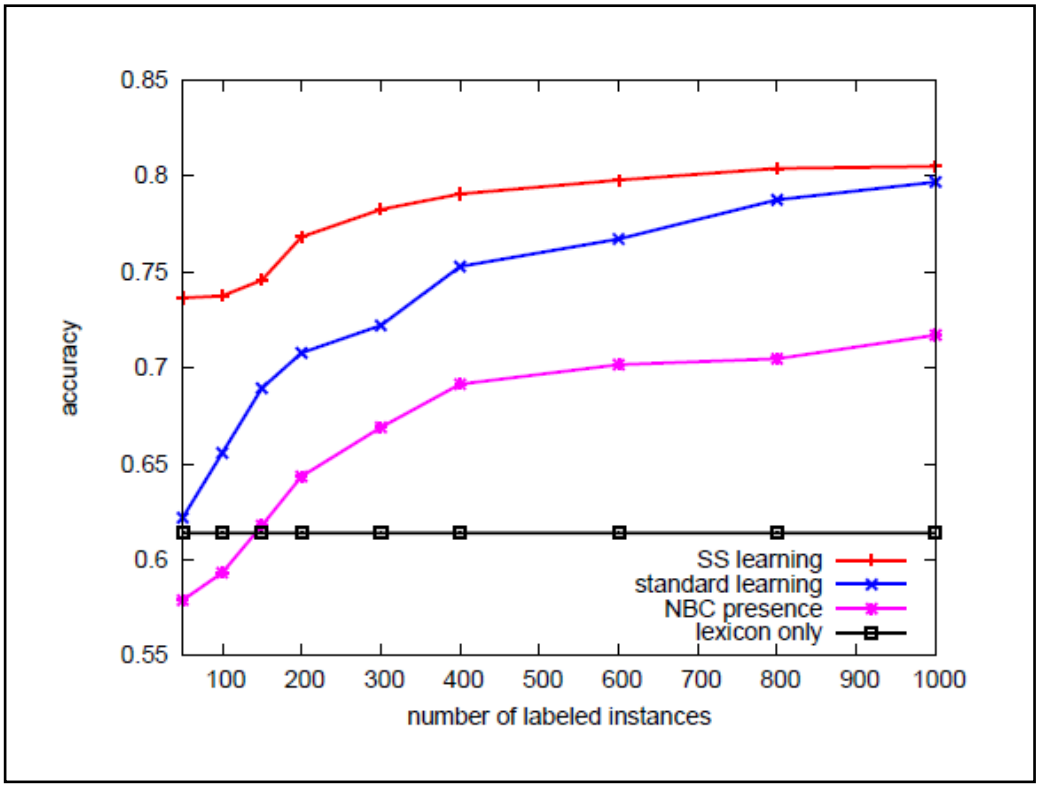
where the L_{ij} are matrix blocks of L of appropriate dimension.

2. Solve above for $c_{aug} = [\ d_{est}^T \quad c^T\ ]^T$ (e.g., using Conjugate Gradient).

3. Estimate sentiment orientation of any document x of interest as:
   orient = sign($c^T x$).

Semi-Supervised Learning

Sample results: sentiment of online movie reviews (IMDB)



Agile Learning

Background

- Basic idea: we wish to develop an algorithm which learns c_{aug}, and thus c, without requiring *any* labeled documents – thereby substantially increasing the agility of the method.
- Naïve approach: solve the optimization problem

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

Unfortunately this formulation performs poorly, basically because with very little labeled data the optimization produces many isolated like-polarity clusters surrounding labeled instances on G_b, resulting in "over-fitted" solutions with little power for generalization.

Agile Learning

Algorithm AL

- Trick: replace L (or L_n) with a power L_n^k, which "smoothes" the polarity estimates assigned to G_b and so reduces over-fitting.
- Mathematics: solve optimization

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

- Algorithm: Solve the set of linear equations

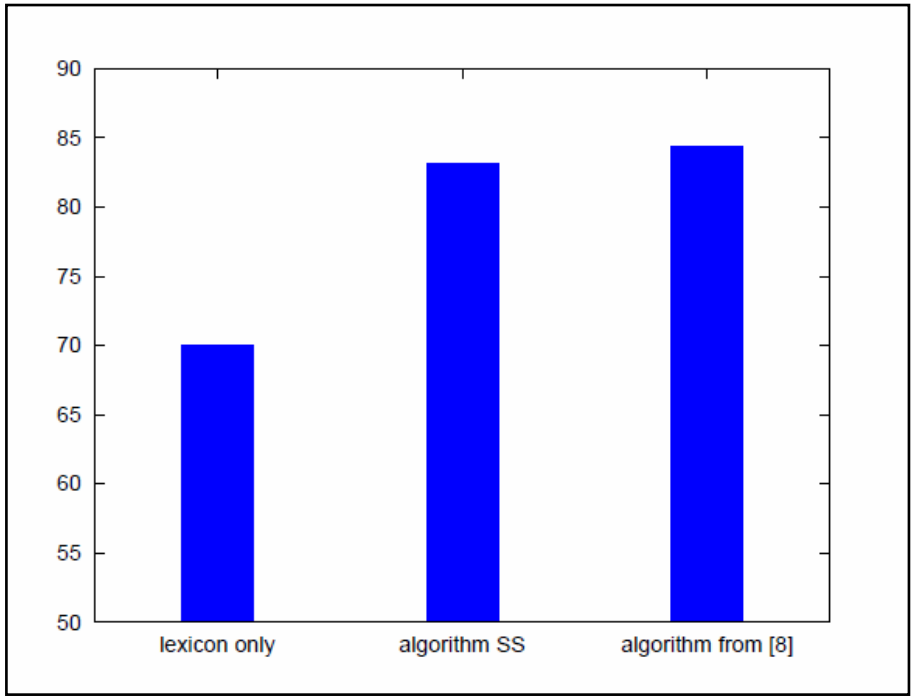
$$\begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} + \beta I_{|V_l|} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix} c_{aug} = \begin{bmatrix} 0 \\ \beta w \\ 0 \end{bmatrix}$$

for $c_{aug} = [\, d_{est}^T \quad c^T \,]^T$ and estimate sentiment via orient = sign($c^T x$).

Agile Learning

Sample results: sentiment of online product reviews (Amazon)

We compare Algorithm AL with the SCL algorithm [Blitzer et al. 2007] trained with 1600 labeled documents; the latter is the "gold standard" for this task.



Sentiment Estimation

method	accuracy
lexicon-only	70.8%
"gold standard"	84.4%
algorithm AL	83.1%

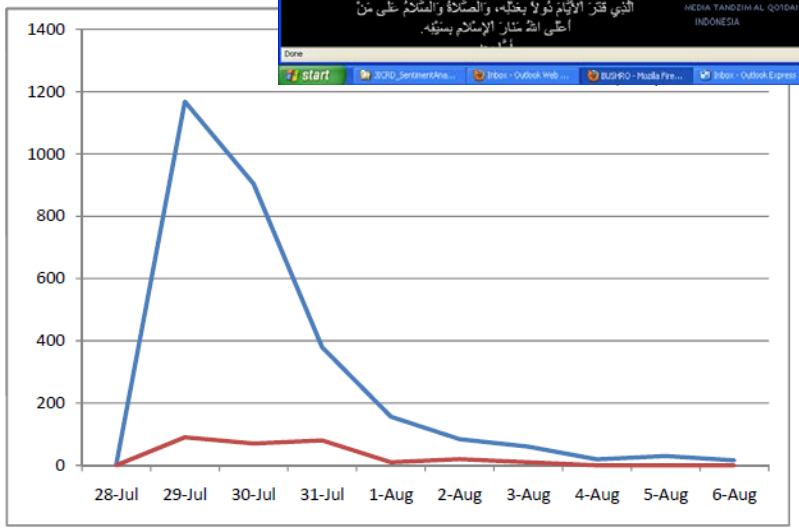
Sentiment Proportion

method	accuracy
algorithm AL	93.3%

National Security Examples

Public sentiment: example one

- Problem: characterize Indonesian public opinion about 26 July 2009 blog post, allegedly by NM Top (www.mediaislam-bushro.blogspot.com), which claimed responsibility for 19 July Jakarta hotel bombings.
- Sample results: application of Algorithm SS to 1.) ~3000 Indonesian language comments made to above blog and 2.) ~500 relevant posts made to other blogs, reveals online Indonesian public reaction was overwhelmingly negative.



National Security Examples

Public sentiment: example two

- Problem: estimate *regional* public opinion regarding former Egyptian President Hosni Mubarak during the weeks prior to the protests that began on 25 January 2011.
- Sample results: application of Algorithm SS to 1.) 100 Arabic blog posts, 2.) 100 Indonesian posts, and 3.) 100 Danish posts reveals:
  - online public opinion regarding Mubarak was largely negative;
  - fraction of negative posts varied by post language, and thus possibly by geographic region.