SAND2011-4696 C
SAND2011-4696C

Estimating Sentiment in Social Media

Kristin Glass* **Rich Colbaugh*†**

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

July 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*.

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.

Introduction

Outline

- Problem formulation:
  - text classification;
  - bipartite graph data model.
- Method One:
  - semi-supervised learning;
  - sample results.
- Method Two:
  - transfer learning;
  - sample results.
- National security applications.

Problem Formulation

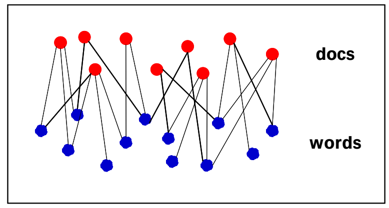
Sentiment analysis as text classification

- 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 and supplement limited labeled data with *auxiliary* information, particularly that which is readily available in online applications.

Problem Formulation

Bipartite graph data model

- Assume given:
  - corpus of n documents, of which $n_l << n$ are labeled ($d \in \Re^{nl}$);
  - modest lexicon V_l of sentiment-laden words ($w \in \Re^{|Vl|}$);
  - corpus of n_S documents from a domain related to target domain.

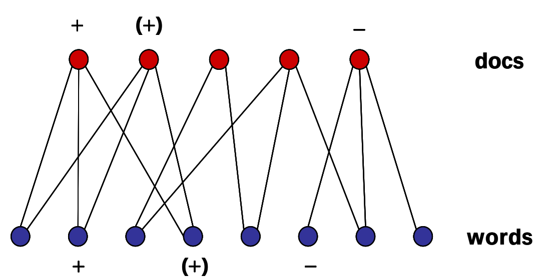


- Analytic approach: leverage information in *unlabeled* documents and/ or *related* documents by:
  - modeling data as a bipartite graph G_b of documents and words;
  - assuming that, in G_b, positive/negative documents (words) will tend to be connected to positive/negative words (documents).

Semi-supervised Learning

Algorithm SS

- 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: implementation

1. Construct the following set of linear equations:

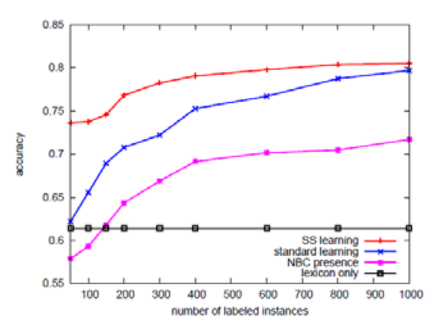
$$\begin{bmatrix} L_{11}+\beta_1 I_{nl} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33}+\beta_2 I_{|V_l|} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix} c_{aug} = \begin{bmatrix} \beta_1 d \\ 0 \\ \beta_2 w \\ 0 \end{bmatrix}$$

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 = $\mathrm{sign}(c^T x)$.

Semi-supervised Learning

Sample results: sentiment of online movie reviews (IMDB)



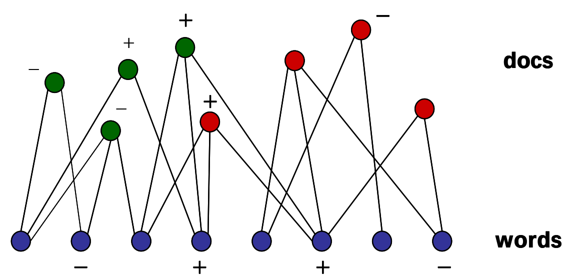
Sentiment Proportion

#Labeled Docs	Accuracy
50	82.7%
200	88.6%
1000	94.5%

Transfer Learning

Algorithm TL

- Basic idea:



- Mathematics:

$$\min_{c_{aug}} \; c_{aug}^T L c_{aug} + \beta_1 \left\| d_{S,est} - k_S d_S \right\|^2 + \beta_2 \left\| d_{T,est} - k_T d_T \right\|^2 + \beta_3 \left\| c - w \right\|^2$$

Transfer Learning

Algorithm TL: implementation

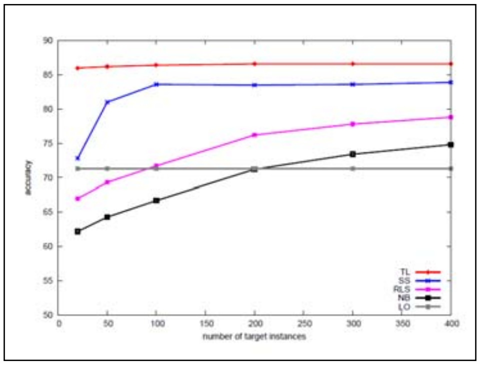
1. Construct the following set of linear equations:

$$\begin{bmatrix} L_{11} + \beta_1 I_{nS} & L_{12} & L_{13} \\ L_{21} & L_{22} + \beta_2 I_{nT} & L_{23} \\ L_{31} & L_{32} & L_{33} + \beta_3 I_{|V_1|} \end{bmatrix} c_{aug} = \begin{bmatrix} \beta_1 k_S d_S \\ \beta_2 k_T d_T \\ \beta_3 w \end{bmatrix}$$

2. Solve above for $c_{aug} = [\; d^T_{S,est} \quad d^T_{T,est} \quad c^T \;]^T$ (e.g., using Conjugate Gradient).
3. Estimate sentiment orientation of any document x of interest as: orient = $\mathrm{sign}(c^T x)$.

Transfer Learning

Sample results: sentiment of online product reviews (Amazon)



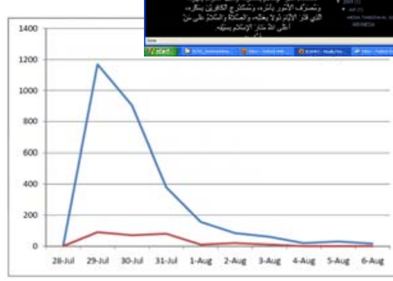
Sentiment Proportion

#Labeled Docs	Accuracy
50	90.7%
100	92.7%
200	94.4%

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 2010.
- 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.