Sentiment Classification Based on Supervised Latent n-gram Analysis

presented by Dmitriy Bespalov

(D. Bespalov, B. Bai, Y. Qi, A. Shokoufandeh (2011) "Sentiment Classification Based on Supervised Latent n-gram Analysis",the *20th ACM Conference on Information and Knowledge Management*)

Drexel
UNIVERSITY

NEC

Overview

- Present three variants of sentiment analysis (SA) tasks
- Bag-of-Words (BoW) representation and its extension for phrases
- Formulate the three SA tasks as machine learning problems
- Present the proposed method for latent representation of n-grams
- Experimental results:
    - Three sentiment analysis tasks
    - Two large-scale datasets: Amazon & TripAdvisor

Definition of Opinion Mining

- Sentiment Analysis (SA) – is the task of extracting subjective information from natural language text
- Consider three variants of SA task
  - **Binary SA**
    - Estimates overall sentiment of text as positive or negative
  - **Multi-scale SA**
    - Determines overall sentiment of text using Likert scale
    - e.g., 5-star system for online reviews
  - **Pair-wise SA**
    - Orders pairs of texts based on sentiment strength and polarity
    - e.g., “very good” vs “good” vs “terrible” vs “absolutely terrible”

Automatic Sentiment Analysis

- Automatic SA can be tackled with machine learning
- Labeled training data is used to bias system's output
- Formulate the three SA tasks:
  - **Binary SA** as a binary classification problem
  - **Multi-scale SA** as ordinal classification
    - preferred for labels that admit ordering
  - **Pair-wise SA** as margin ranking loss optimization

Prior Work on Automatic SA

- Prior work primarily considered binary SA task
- In-depth survey of the automatic SA methods are presented in [1]
- Unsupervised extraction of aspect-sentiment relationship was used in [2]
- Modeling content structure in semi-supervised manner was shown to improve SA accuracy in [3]
- Discriminative (fully) supervised methods result in current state-of-art
    - Linear SVM trained on BoW representation [4]

[1] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2008*.

[2] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. *HLT 2010.*

[3] C. Sauper, A. Haghighi, and R. Barzilay. Incorporating content structure into text analysis applications. *EMNLP 2010.*

[4] G. Paltoglou and M. Thelwall. A study of information retrieval weighting schemes for sentiment analysis. *ACL 2010*.

Bag-of-Words Representation for Text

“Think and wonder, wonder and think.”

and	2
think	2
wonder	2

- **Bag-of-Words (BoW)** model treats text as order-invariant collection of features
  - Enumerate all unique words in text corpus and place into dictionary $\mathcal{D}$
  - Let $\mathbf{x} = (w_1, \cdot \ \ , w_N)$ denote a document from corpus
  - Define canonical basis vector with single non-zero entry at position w_i :

$$\mathbf{e}_{w_i} = (0, \ldots, 1, \ldots, 0)^\top$$

  - Define BoW representation of document $\mathbf{x}$:

$$\tilde{\mathbf{e}}_{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_{w_i} \qquad \dim(\tilde{\mathbf{e}}_{\mathbf{x}}) = \dim(\mathbf{e}_{w_i}) = |\mathcal{D}| \times 1$$

  - Optionally, assign weights (e.g., TF-IDF, BM25) to every word

Word Phrases in Opinion Mining

- Short phrases better capture sentiment than words
  - Consider words "recommend" and "book"

"I absolutely recommend this book"

"I highly recommend this book"

"I recommend this book"

"I somewhat recommend this book"

"I don't recommend this book"

Modeling Phrases in BoW Model

"the film is palpable evil genius"

"the film"

"film is"

"is palpable"

"palpable evil"

"evil genius"

- BoW extension to encode **positional information**
  - Collect all phrases with n words or less (i.e., n-grams) from corpus
  - Add n-grams to set Γ and use them as features in BoW model:

$$\dim(\mathbf{e}_{w_i}) = |\Gamma| \times 1, \quad |\Gamma| = O(|\mathcal{D}|^n$$

The Proposed Method

- Adopt method by Collobert and Weston [5] to SA tasks
- Embed all sliding n-gram windows from text in latent space
- Compute latent text representation as centroid of all n-grams
- The process is modeled as a feed-forward neural network
- Parameters for latent projections and classifiers are jointly estimated via backpropagation using stochastic gradient descent

[5] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. *ICML 2008.*

Feed-forward Deep Architectures

Input Vector $\mathbf{x}$

Linear Projection $\mathbf{p}_1 = \mathbf{A}\mathbf{x} + \mathbf{b}$

Non-linear Transfer Function $\mathbf{p}_2 = \mathrm{h}(\mathbf{p}_1)$

$$\tanh(t) = \frac{e^{2t} - 1}{e^{2t} + 1} \qquad \mathrm{sigmoid}(t) = \frac{1}{1 + e^{-t}}$$

…

Optimization objective (i.e., NN Criterion) $\mathrm{L}(\mathbf{p}_\mathbf{k})$

Backpropagation & Stochastic Gradient Descent

- **Backpropagation** [6] is supervised learning method for **neural network (NN)**
  - Using backward recurrence it jointly optimizes all NN parameters
  - Requires all activation functions to be differentiable
  - Enables flexible design in deep NN architecture
  - Gradient descent is used to (locally) minimize objective:

$$\mathbf{A}^{\mathbf{k}+1} = \mathbf{A}^{\mathbf{k}} - \eta \frac{\partial \mathrm{L}}{\partial \mathbf{A}^{\mathbf{k}}}$$

- **Stochastic Gradient Descent (SGD)** [7] is first-order iterative optimization
  - SGD is an **online learning** method
  - Approximates “true” gradient with a gradient at one data point
  - Attractive because of low computation requirement
  - Rivals **batch learning** (e.g., SVM) methods on large datasets

[6] Y. LeCun et al. 1998. Efficient BackProp.

[7] Olivier Bousquet and Ulrike von Luxburg. 2004. Stochastic Learning.

Latent Embedding for BoW Document

BoW Document $\mathbf{X}$

Sparse Linear $\mathbf{d}_{\mathbf{x}} = \mathbf{G}\,\tilde{\mathbf{e}}_{\mathbf{x}}$

SA Task $g\left(\mathbf{d}_{\mathbf{x}}\right)$

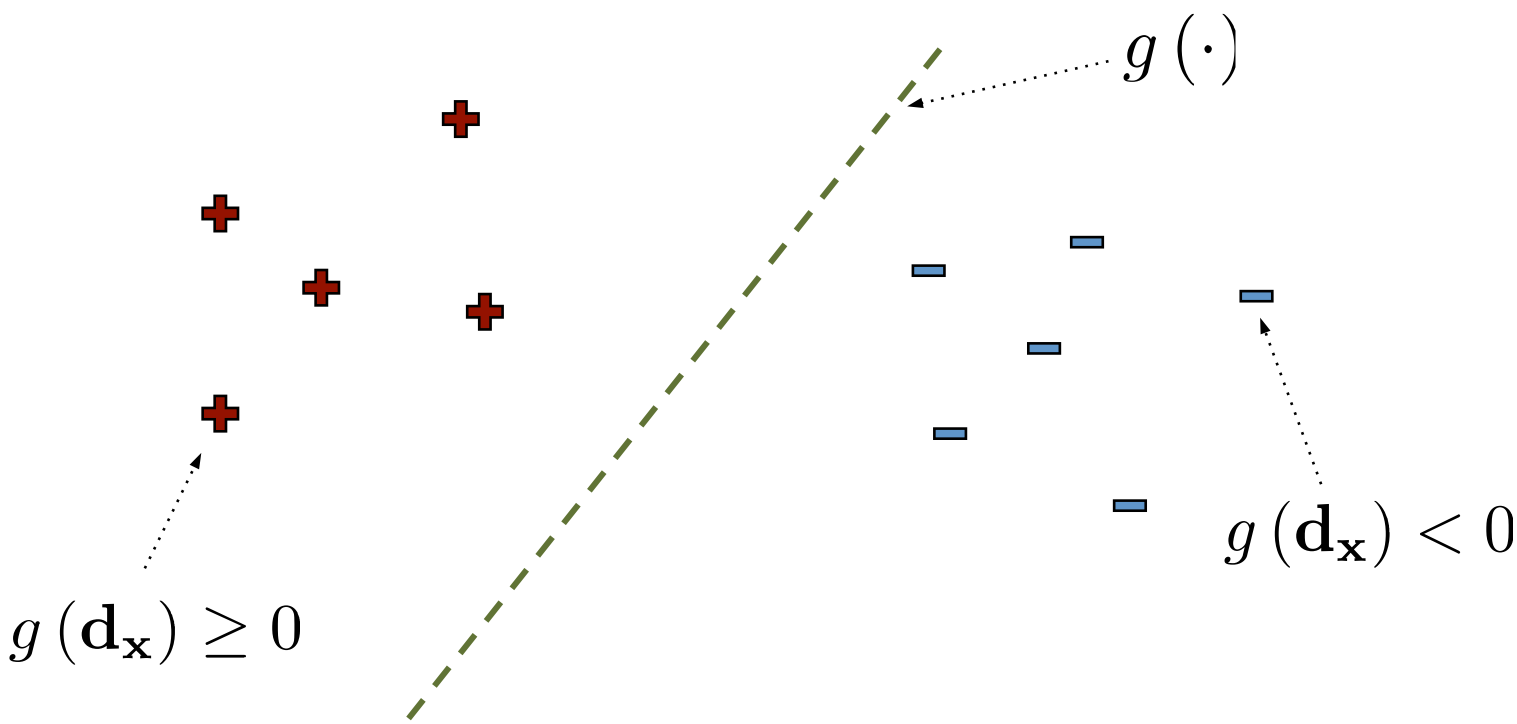
$$\tilde{\mathbf{e}}_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_{w_i}$$

$$\mathbf{e}_{w_i} = (0, \ldots, 1, \ldots, 0)^{\top}$$

$$\dim(\mathbf{e}_{w_i}) = |\Gamma| \times 1, \quad |\Gamma| = O(|\mathcal{D}|^{n}$$

Binary Sentiment Analysis

- Formulate binary SA task as classification
  - Typical formulation for Support Vector Machines (SVM)



SVM Classification [8] for BoW Document

BoW Document $\mathbf{X}$

SVM Classifier $g\left(\mathbf{d}_{\mathbf{x}}\right) = \mathbf{G}\,\tilde{\mathbf{e}}_{\mathbf{x}}$

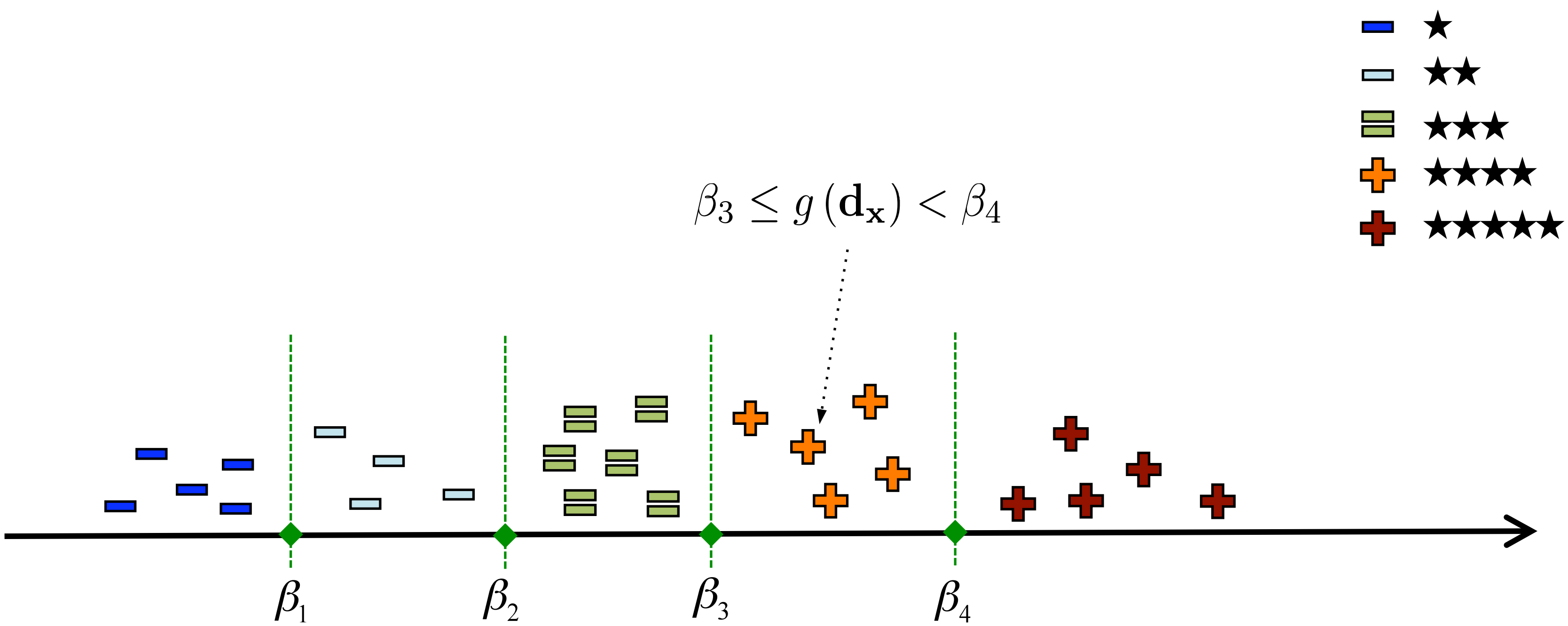
$$\dim(\mathbf{G}) = 1 \times |\Gamma|$$

SVM adds regularization term to objective: $\frac{1}{2}\left\|\mathbf{G}\right\|_2^2$

[8] C. Cortes and V. Vapnik, Support-Vector Networks, *Machine Learning, 20, 1995.*

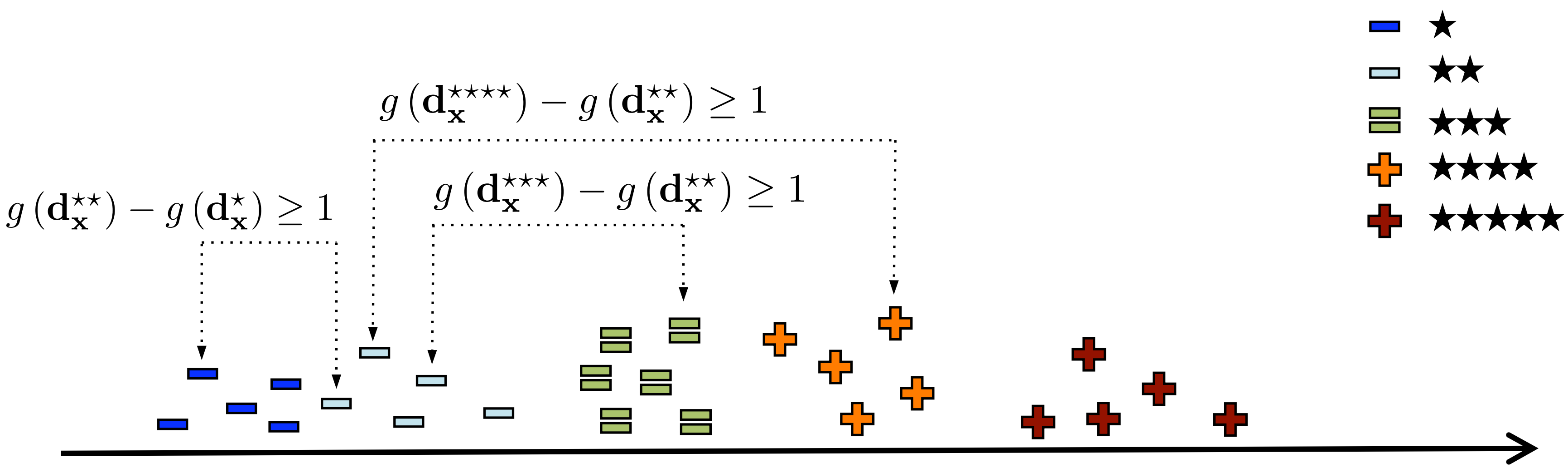
Multi-Scale Sentiment Analysis

- Formulate multi-scale SA as ordinal classification
  - Only perceptron-based models are considered

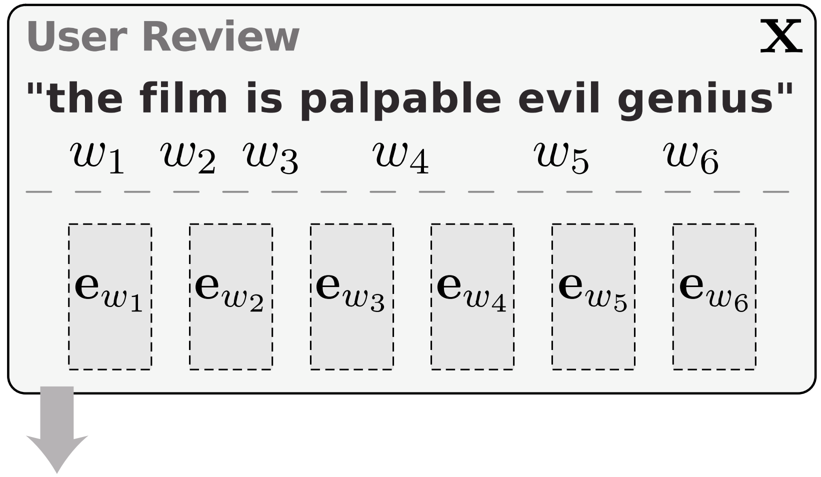


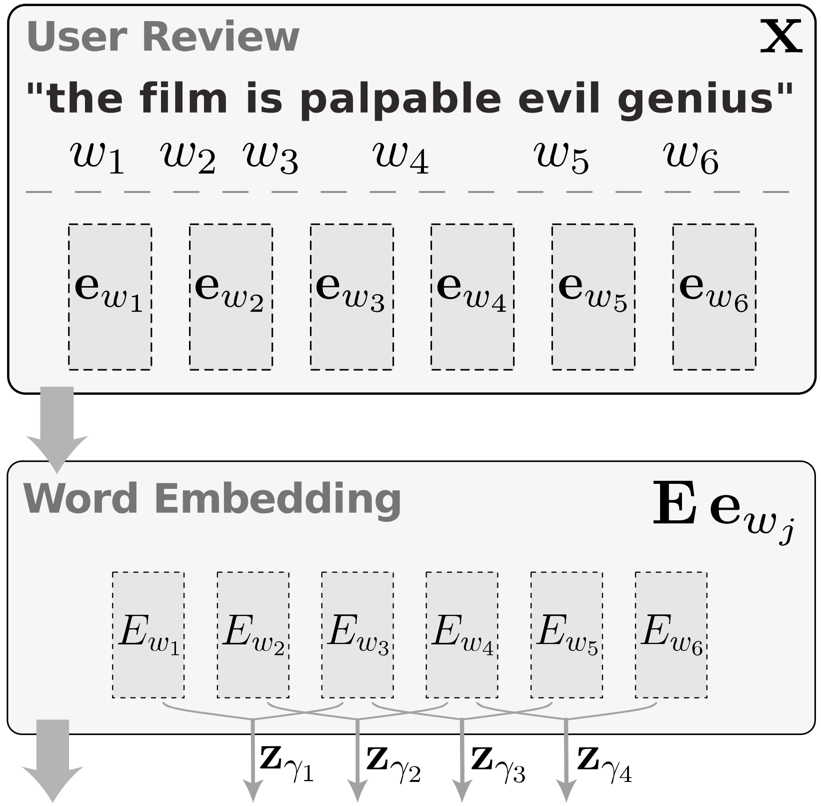
Pair-Wise Sentiment Analysis

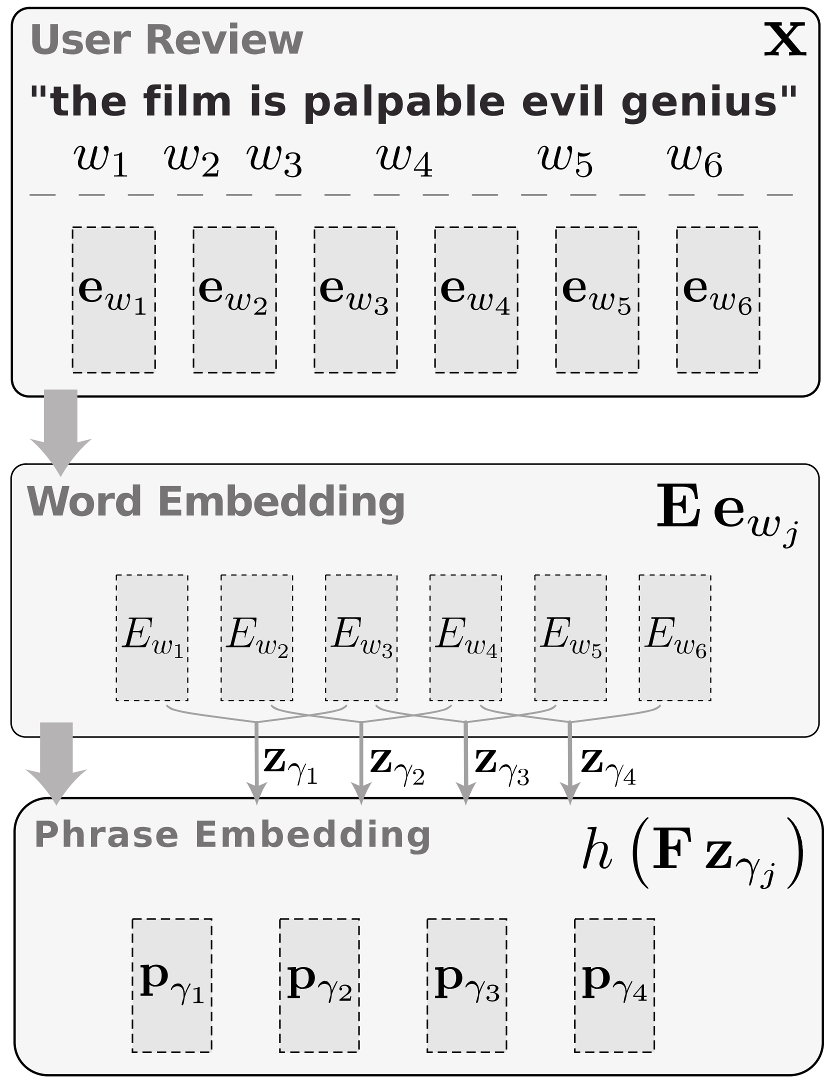
- Formulate pair-wise SA as margin ranking loss
  - No labels are required
  - Use labels for pair-wise supervised signal
  - Only perceptron models are considered



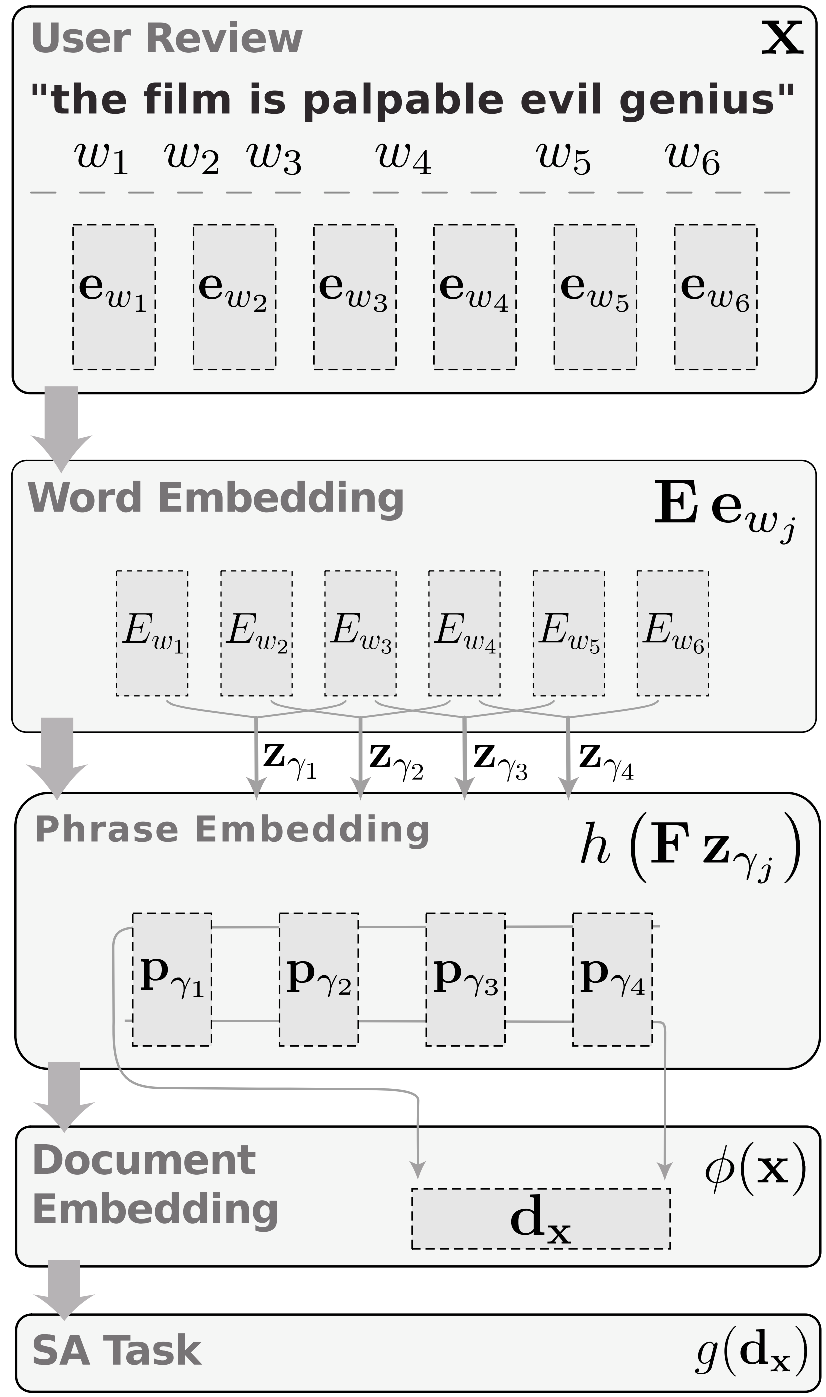
Latent Embedding of n-grams (SLNA)


User Review
X
"the film is palpable evil genius"
w_1 w_2 w_3 w_4 w_5 w_6
$\mathbf{e}_{w_1}$ $\mathbf{e}_{w_2}$ $\mathbf{e}_{w_3}$ $\mathbf{e}_{w_4}$ $\mathbf{e}_{w_5}$ $\mathbf{e}_{w_6}$


User Review
$\mathbf{X}$
"the film is palpable evil genius"
w_1 w_2 w_3 w_4 w_5 w_6
$\mathbf{e}_{w_1}$ $\mathbf{e}_{w_2}$ $\mathbf{e}_{w_3}$ $\mathbf{e}_{w_4}$ $\mathbf{e}_{w_5}$ $\mathbf{e}_{w_6}$
Word Embedding
$\mathbf{E}\,\mathbf{e}_{w_j}$
E_{w_1} E_{w_2} E_{w_3} E_{w_4} E_{w_5} E_{w_6}
$\mathbf{z}_{\gamma_1}$ $\mathbf{z}_{\gamma_2}$ $\mathbf{z}_{\gamma_3}$ $\mathbf{z}_{\gamma_4}$



User Review
$\mathbf{X}$
"the film is palpable evil genius"
w_1 w_2 w_3 w_4 w_5 w_6
$\mathbf{e}_{w_1}$ $\mathbf{e}_{w_2}$ $\mathbf{e}_{w_3}$ $\mathbf{e}_{w_4}$ $\mathbf{e}_{w_5}$ $\mathbf{e}_{w_6}$
Word Embedding
$\mathbf{E}\,\mathbf{e}_{w_j}$
E_{w_1} E_{w_2} E_{w_3} E_{w_4} E_{w_5} E_{w_6}
$\mathbf{z}_{\gamma_1}$ $\mathbf{z}_{\gamma_2}$ $\mathbf{z}_{\gamma_3}$ $\mathbf{z}_{\gamma_4}$
Phrase Embedding
$h\left(\mathbf{F}\,\mathbf{z}_{\gamma_j}\right)$
$\mathbf{p}_{\gamma_1}$ $\mathbf{p}_{\gamma_2}$ $\mathbf{p}_{\gamma_3}$ $\mathbf{p}_{\gamma_4}$


User Review
$\mathbf{x}$
"the film is palpable evil genius"
w_1 w_2 w_3 w_4 w_5 w_6
$\mathbf{e}_{w_1}$ $\mathbf{e}_{w_2}$ $\mathbf{e}_{w_3}$ $\mathbf{e}_{w_4}$ $\mathbf{e}_{w_5}$ $\mathbf{e}_{w_6}$
Word Embedding
$\mathbf{E}\,\mathbf{e}_{w_j}$
E_{w_1} E_{w_2} E_{w_3} E_{w_4} E_{w_5} E_{w_6}
$\mathbf{z}_{\gamma_1}$ $\mathbf{z}_{\gamma_2}$ $\mathbf{z}_{\gamma_3}$ $\mathbf{z}_{\gamma_4}$
Phrase Embedding
$h\left(\mathbf{F}\,\mathbf{z}_{\gamma_j}\right)$
$\mathbf{p}_{\gamma_1}$ $\mathbf{p}_{\gamma_2}$ $\mathbf{p}_{\gamma_3}$ $\mathbf{p}_{\gamma_4}$
Document Embedding
$\phi(\mathbf{x})$
$\mathbf{d}_\mathbf{x}$
SA Task
$g(\mathbf{d}_\mathbf{x})$

Advantages of Proposed Method

- Augmenting BoW representation with n-grams results in exponentially exploding feature space
- The proposed method requires only uni-gram features
- Each uni-gram feature contributes to all latent n-grams that contain the feature
- Parameter space for our model grows linear with size of n-gram, recognized by the model
- The proposed method results in the state-of-art performance for three SA tasks

Experimental Results: Datasets

- Use two large-scale sentiment datasets
  - Amazon & TripAdvisor
- Amazon contains product reviews from 25 categories
  - e.g., apparel, automotive, baby, DVDs, electronics, magazines
- TripAdvisor contains hotel reviews from across the globe
  - Consider only overall ratings for the reviews
- Create balanced 70/30% train-test split
- Sample training set to obtain validating set
- Limit vocabulary to terms with highest mutual information (MI) shared with the binary labels [9]

[9] J. Blitzer et al. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. *ACL 2007.*

Datasets (cont'd)

	Amazon	**TripAdvisor**
$\star$	103,953	15,152
$\star\star$	80,278	20,040
$\star\star\star$	0	25,968
$\star\star\star\star$	48,086	15,141
$\star\star\star\star\star$	136,145	20,051
Train	237,900	64,445
Test	110,562	28,907
Validate	20,000	3,000
Total	368,462	96,352
$\lvert\mathcal{D}\rvert$	448,146	158,997
$\lvert\Gamma_1\rvert$	54,334	54,000
$\lvert\Gamma_2\rvert$	127,337	127,000

Both dataset splits are available at: http://mst.cs.drexel.edu/datasets

Experimental Results: Binary SA

Method	Amazon	TripAdvisor
BOW Prc 1g	10.96	8.27
BOW Prc 2g	7.59	**7.37**
BOW SVM 1g	11.10	8.89
BOW SVM 2g	7.45	7.47
BOW SVM Δ-IDF 1g	10.91	8.74
BOW SVM Δ-IDF 2g	7.39	7.96
BOW LSI SVM 1g	21.40	24.18
SLNA	9.84	8.92
SLNA LSI	**7.12**	8.33
SLNA LT-FIX	15.4	-

Proposed method, based on latent n-gram embedding.

[10] G. Paltoglou and M. Thelwall. A study of information retrieval weighting schemes for sentiment analysis. *ACL 2010*.

Experimental Results: Pair-wise SA

Method	**Amazon**	**TripAdvisor**
BOW Prc 2g	12.5	16.0
SLNA LSI	**10.2**	**13.7**

Experimental Results: Multi-class SA

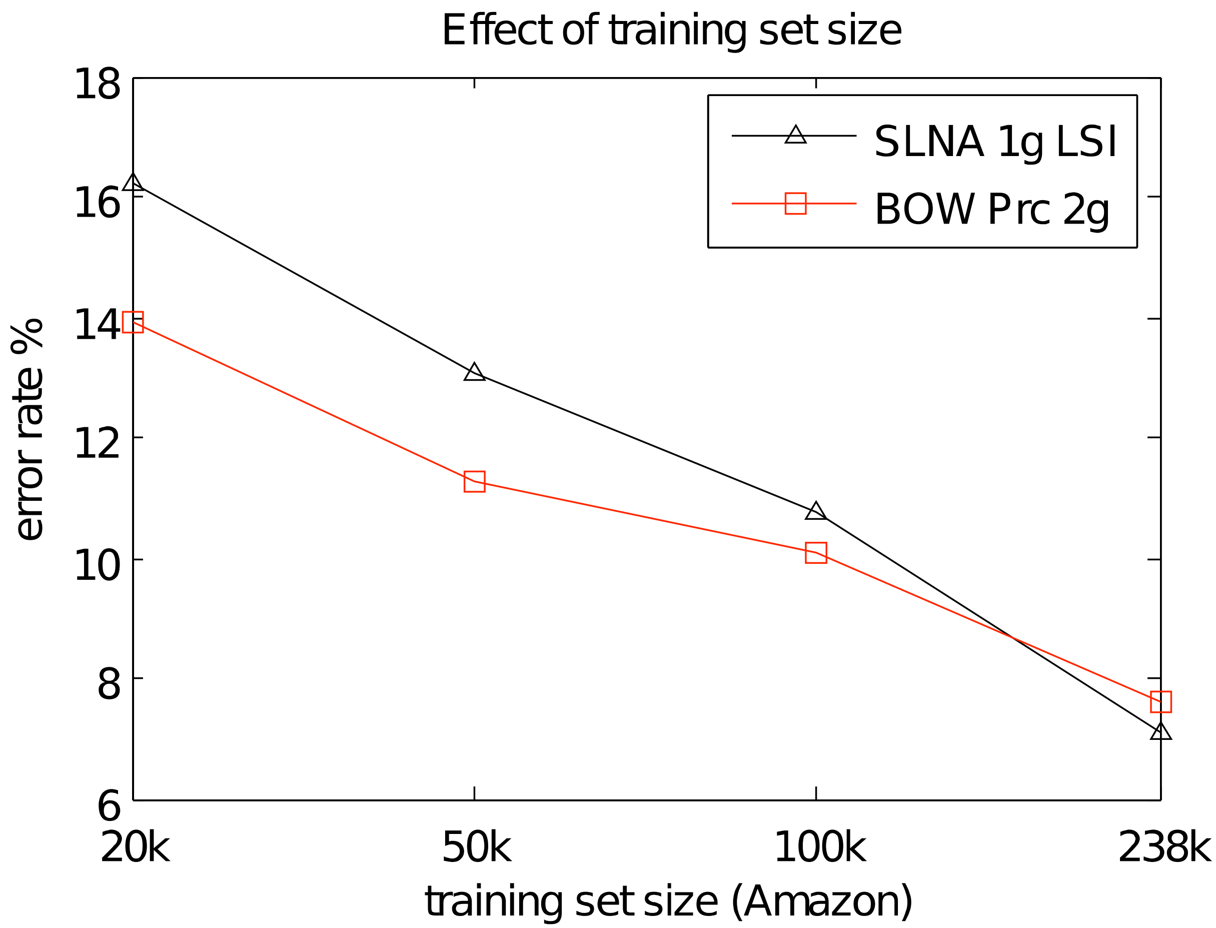
Method	Amazon	TripAdvisor
BOW Prc 2g	37.8	44.9 (52.1)
BOW Prc 2g RL	35.8	41.5 (51.6)
SLNA LSI	30.7	42.7 (51.4)
SLNA LSI RL	**28.2**	**39.6 (49.2)**

Both models are initialized with parameters pre-trained on pair-wise SA task

- Using **SLNA LSI RL** model we obtained MSE = 1.03 on Amazon dataset.
- Previously reported MSE ≈ 1.5 on a small subset of Amazon dataset [11]

[11] Y. Mansour et al. Domain adaptation with multiple sources. *NIPS 2008*.

Experimental Results: Training Set Size



Q&A