

# Discovering Voter Preferences in Blogs using Mixtures of Topic Models

Pradipto Das, Rohini Srihari, Smruthi Mukund

SUNY Buffalo, USA

AND Conference, July 23-24, 2009



# Part I

## Motivation

# Political Speeches

- Sample Speech from Obama

*"Provide \$50 billion to Jumpstart the Economy and Prevent 1 Million Americans from Losing Their Jobs: This relief would include a \$25 billion State Growth Fund to prevent state and local cuts in health, education, housing, and heating assistance or counterproductive increases in property taxes, tolls or fees. The Obama-Biden relief plan will also include \$25 billion in a Jobs and Growth Fund to prevent cutbacks in road and bridge maintenance and fund school repair - all to save more than 1 million jobs in danger of being cut."*

- Sample Speech from McCain

*"The Arizona senators Pension and Family Security Plan would eliminate taxes on unemployment benefits for those making less than \$100,000 annually. ... Additionally, the plan would halve the capital gains tax on stock profits to 7.5 percent for two years. Its a plan that would broadly benefit the economy and would put cash into the hands of those damaged by the stock market, McCain economic adviser Douglas Holtz-Eakin said in a conference call with reporters. It would stabilize some of the financial insecurity we see."*

# Blogs

- Sample Blog Post

*"[McCain speaking in front of the NRA in May, 2008] John McCain's campaign won't say whether he's for or against allowing suspected terrorists to buy guns, as he tries to pander to his lobbyist pals and the Republican pro-gun base but wanders into the War On Some Terror minefield by mistake. Sen. John McCain portrays himself as a strong supporter of Second Amendment rights. But does that extend to gun rights for suspected terrorists? His campaign won't say where he stands on a bill to eliminate a gun-control loophole that even the Bush administration wants closed: a gap in federal law that inhibits the government from stopping people on terrorist watch lists from buying guns."*

## Problem Statement

Can you identify towards which election candidate is the blogger inclined and more importantly why?

# The hard question - “why?”

- The goal here is not really to classify blogs as pro- $Cand_1$  or pro- $Cand_2$  rather to discover as to why a blogger is inclined towards a particular candidate
- Helps answer questions like - “does he like the candidate for his charisma or does he like the candidate based on what he likes to achieve?”

## Supervised vs. Unsupervised learning perspective

- classification → typically don't need to know “what” is causing the phenomenon, rather explain the phenomenon in terms of hand-coded features
  - Class conditional densities generate the decision function to demarcate the dichotomy*
- Clustering → assume something about what is causing the phenomenon and group similar things together
- “Topics” from topic models help us group words into themes

The blogs are not labeled and unsupervised learning seems like a natural choice

# Theme it up!

## Theme Season sale! Buy 2, get 1 free

- topic models explore the generating process of the latent themes or topics in a document
- Each word position in a document is filled up by first choosing a “topic” for the document and then choosing a word “label” for that topic
- A “topic”  $\rightarrow$  distribution over words
  - $\rightsquigarrow$  A “document”  $\rightarrow$  distribution over such topics
- run a  $K$  (model complexity parameter) topic model over two ground truth speeches fix  $K$  based on inference on likelihoods of the ground truth corpus
  - $\Rightarrow$  collect sets of model parameters (topic-word probabilities) one for each candidate
- run the same  $K$  parameter topic model on the blogs (for free!)
  - $\Rightarrow$  collect inference on latent variables (word-topic probabilities) per blog document

## Part II

### The method

# Base and sentiment-weighted base models

## Base topic model (CTM)

- Find  $p(w_j | z_j, \beta)$   
the  $z_j$ 's are indicators of random variables having a logistic normal prior

## Sentiment weighted base topic model

- Weight the base model by topic sentiment
- Find  $p(\textit{Sentiment} = \textit{"positive"}, w_{j_{blog}} | z_j, \beta)$  for the positive sentiment weighted base topic model i.e. evaluate the following for the 3 sentiment models (positive, negative and objective)

$$\bullet p(s_{\textit{SentimentType}}, w_{j_{blog}} | z_j, \beta) \propto p(w_{j_{blog}} | z_j, \beta) \delta(w_{j_{blog}}, w_{x_{cand}})$$

$$\times (p(z_j | s_{\textit{SentimentType}}) \times p_{\textit{prior}}(s_{\textit{SentimentType}}))$$

# One way of hard-labeling blogs

## KL Divergence

- We can evaluate each of the 4 models on *each* word of (blog, candidate speech) pair
- Then for each blog document,
  - we compare the divergences of each word in the blog to that word in the candidate speeches (only if the word is found in both the speeches)
  - if majority of the words in the blog have lower topic divergence for  $candidate_1$  ground truth, then the blog is labeled to be more inclined to  $candidate_1$
- The most probable words under the inferred topics of the blog words and those of  $candidate_1$ 's ground truth are treated as reasons for the inclinations.

# Combining Base and weighted-base models

Model combination using collaborative objective function

$$E_i = \frac{\alpha}{2} \sum_{i=1}^M (t_i - f(\mathbf{x}_i^T \mathbf{w}_j))^2 - \frac{\beta}{2} \sum_{j=1}^H \sum_{l=1, l \neq j}^H \cos^2(\mathbf{w}_j^T \mathbf{w}_l)$$

where,

$$f(\mathbf{x}_i^T \mathbf{w}_j) = \frac{1}{(1 + \exp(-\mathbf{x}_i^T \mathbf{w}_j))}$$

- Adapt the EM algorithm for expert voting (Note: the  $t_i$ 's are not ground truth blog labels, rather labels assigned by the models)

We note that there are 4 models - so  $h \in \{1, 2, 3, 4\}$  and  $j, l \in \mathbf{h}$

- each blog document  $i$  under model  $j$  is now a triplet  $(t_i, \mathbf{x}_i, \mathbf{w}_j)$
- model  $h = 1$  is the base topic model (CTM),  $h = 2$  is CTM weighted with positive sentiments etc.
- The combined divergence for each word in the blog vocabulary and for each model  $h$  is given by

$$\mathbf{w}_j^{(v)} = \frac{M}{\sum_{i=1}^M \left( \frac{1}{1+w_{j1}^{(v)}} + \dots + \frac{1}{1+w_{jM}^{(v)}} \right)}$$

where  $M$  is the number of blog documents and  $w_{(.)}^{(v)}$  is the weight of  $v^{th}$  word in the blog vocabulary in blog  $i$

# E-Step for model combination

## Estimate expert distribution for each datum

Introduce a set of variational distributions  $Q = \{Q_i(h_i)\}_{i=1}^M$  over the hidden indicators for each of the data points (i.e. blog)

$$\begin{aligned} \mathcal{L}(\Theta_{\mathbf{w}}) &\equiv \log P(\{\mathbf{x}_i, t_i\}_{i=1}^M | \Theta_{\mathbf{w}}) \\ &\geq \sum_{i=1}^M \sum_{j=1}^H Q_i(h_i) \log \left( \frac{P(h_i) P(\mathbf{x}_i, t_i | \mathbf{w}_j, h_i)}{Q_i(h_i)} \right) \equiv \mathcal{F}(\Theta_{\mathbf{w}}, \mathbf{Q}) \end{aligned}$$

## Updating Q

$$\log P(\mathbf{x}_i, t_i, h_i | \Theta_{\mathbf{w}}) - \log Q_i(h_i) + 1 - \lambda_i = 0$$

where  $\lambda_i$  are the  $M$  Lagrange Multipliers, one for the constraint of  $Q$  to be a distribution over experts for each *document*. Solving this equation leads to

$$Q_i(h_i = j) = \frac{\pi_j \times P(\mathbf{x}_i, t_i | \mathbf{w}_j, h_i = j)}{\sum_{l=1}^H \pi_l \times P(\mathbf{x}_i, t_i | \mathbf{w}_l, h_i = j)}$$

# M-Step for model combination

Estimate ML parameter settings for "experts"

$$\frac{\sum_{i=1}^M Q_i(h_i = j)}{\pi_j} + \lambda = 0$$

$$\Rightarrow \pi_j = \frac{\sum_{i=1}^M Q_i(h_i = j)}{M}$$

Estimate ML parameter settings for model weights

$$\frac{\partial \mathcal{F}(\Theta_{\mathbf{w}}, \{Q_i(h_i)\}_{i=1}^M)}{\partial \mathbf{w}_j} = - \sum_{i=1}^M Q_i(h_i = j) \frac{\partial E_{i,j}}{\partial \mathbf{w}_j} + C$$

where  $C$  is some constant. Since the derivative depends on  $\mathbf{w}_l$ 's, we use negative gradient descent to update  $\mathbf{w}_j$  with

$$-\eta \Delta \mathbf{w}_j = \sum_{i=1}^M [\eta \alpha Q_i(h_i = j) (t_i - f(\mathbf{x}_i^T \mathbf{w}_j)) f(\mathbf{x}_i^T \mathbf{w}_j) (1 - f(\mathbf{x}_i^T \mathbf{w}_j)) \mathbf{x}_i -$$

$$\eta \left( \sum_{i=1}^M Q_i(h_i = j) \right) \sum_{l=1, l \neq j}^H \beta \cos(\mathbf{w}_j^T \mathbf{w}_l) \sin(\mathbf{w}_j^T \mathbf{w}_l) \mathbf{w}_l]$$

## Part III

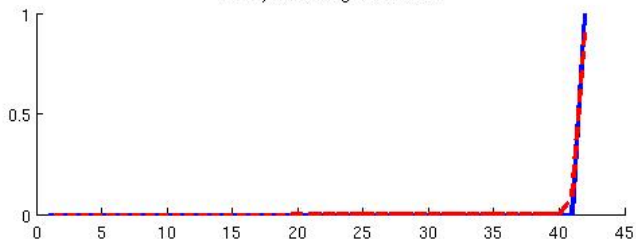
### Examples

# Positive Cases

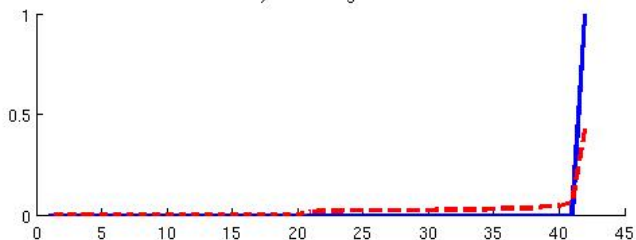
Topic similarities			
Dataset ↓	term in blog	topic in can- didate speech	Few most probable stemmed words un- der the topic
Obama	lobbyist	4	promis america work american time coun- tri chang live care must tax peopl famili democrat economi job
Mccain	lobbyist	23	campaign outrag time press polit voter point charg convent discuss rate emot shift attack
Blog#26	lobbyist	12	wall street polici campaign peopl secur time econom blogger social regul crisi
BlogText	.... as Mccain tries to pander to his lobbyist pals and the Republican pro-gun base but wanders into the War On Some Terror minefield by mistake. ....		

# Graphs for the positive theme matching

"Lobbyist for Blog vs. Obama"



"Lobbyist" for Blog vs. Mccain



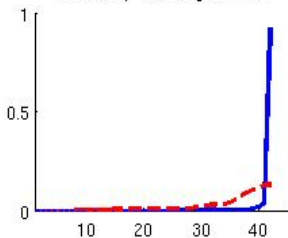
# Negative cases

Topic similarities			
Blog vs. ↓	term in blog	topic in can- didate speech	Few most probable stemmed words un- der the topic
Obama	economy	0	energi oil invest effici feder nation percent technolog fuel build advanc develop
	health	18	plan make cost health american care presid system famili requir
Mccain	economy	24	plan tax econom account save job economi stock pro- pos measur
	health	41	bush administr re- publican support care state issu cam- paign peopl forc talk health
BlogText	Bush-McCain policies have ruined the US economy says this new Obama ad, and now MCain wants to do the same to our health care		

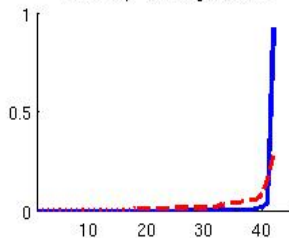
"economy/health" {24} → health econom crisi care polici bank insur market retir today economi decad  
deregul cover claim save current creat magazin

# Graphs for negative cases

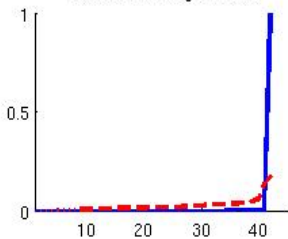
"Economy" for Blog-Obama



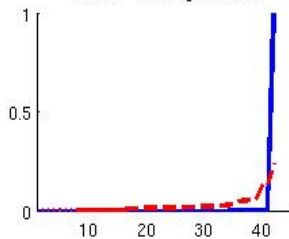
"Economy" for Blog-Mccain



"Health" for Blog-Obama



"Health" for Blog-Mccain



# Sarcasms and jokes

*"Paris Hilton has responded to John McCain for including her in one of his recent campaign ads. To prevent any confusion, please note that, although she does make more sense than John McCain, ... McCain policy loses all substance once you get beyond his attacks on Obama. Paris did make some points but they were fundamentally flawed."*

## Problems with sarcasms

- In real-life blog data we observed that the blog themes were mostly activated by words that McCain used or were used in the reports about him
- Except that the authors were using them to show sarcastic criticisms.

# Some Recall scores

Recall for Blogger Inclinations				
Blogger Inclination	CTM	positive-CTM	negative-CTM	objective-CTM
Obama	0.50	0.54	0.54	0.81
Mccain	0.33	0.33	0.33	0.33

**Note:** Blog classification was never the goal and simple KL divergence of word over topic distributions is not a good measure for the categorization task

## Part IV

The End

That's all folks - Thanks!

**Questions?**