Collective Supervision of Topic Models for Predicting Surveys with Social Media

Michael J. Paul

University of Colorado, Boulder

February 14, 2016

Joint work with Adrian Benton, Braden Hancock, and Mark Dredze

Introduction

- Social media text can be analyzed to understand population-level attributes
  - Public health [1, 4, 6]
  - Political sentiment [5]
- Social media data can augment and complement traditional survey data
  - Advantages: large scale, real time, low cost

Introduction

Two related tasks of interest:

- **Prediction:** estimating survey values for populations from social media features
  - Useful for surveys with limited resources, e.g., gaps in time or geography
- **Analysis:** summarizing public opinions through social media content analysis
  - What text features are correlated with survey values?

Introduction

Two related tasks of interest:

- **Prediction:** estimating survey values for populations from social media features
  - Useful for surveys with limited resources, e.g., gaps in time or geography
- **Analysis:** summarizing public opinions through social media content analysis
  - What text features are correlated with survey values?

Challenge: how to train models that use features at the document level but make predictions at the population level?

Introduction

Two related tasks of interest:

- **Prediction:** estimating survey values for populations from social media features
  - Useful for surveys with limited resources,
    e.g., gaps in time or geography
- **Analysis:** summarizing public opinions through social media content analysis
  - What text features are correlated with survey values?

Challenge: how to train models that use features at the document level but make predictions at the population level?

- **Collective supervision:** supervision is given at the level of a *collection* of documents, rather than individual documents
  - e.g., proportion of population within each US state

Introduction

Topic models can help:

- **Prediction:** estimating survey values for populations from social media features
  - Topic models can learn low-dimensional, generalizable features that can be used in predictive models
- **Analysis:** summarizing public opinions through social media content analysis
  - Topic models are interpretable: we can better understand public opinion if we can see which topics are correlated with surveys

Introduction

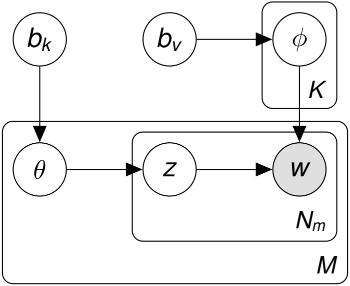
Topic models can help:

- **Prediction:** estimating survey values for populations from social media features
  - Topic models can learn low-dimensional, generalizable features that can be used in predictive models
- **Analysis:** summarizing public opinions through social media content analysis
  - Topic models are interpretable: we can better understand public opinion if we can see which topics are correlated with surveys

Challenge: how to train topic models to learn correlations with surveys?

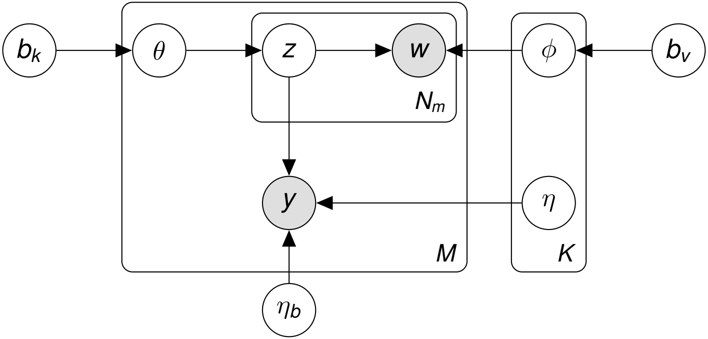
- This talk: modify topic models to incorporate collective supervision
  - We extend different types of topic models in different ways, and compare

Latent Dirichlet Allocation (LDA)



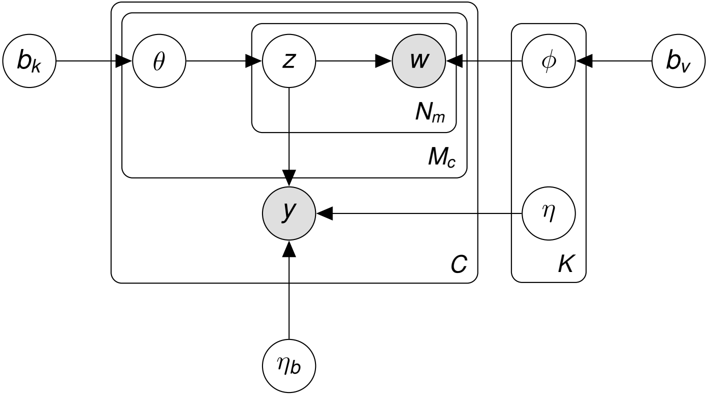
- $\tilde{\theta}_{mk} = \exp(b_k); \theta_m \sim \mathrm{Dirichlet}(\tilde{\theta}_m)$
- $\tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \mathrm{Dirichlet}(\tilde{\phi}_k)$
- $z_{mn} \sim \theta_m; w_{mn} \sim \phi_{z_{mn}}$

Supervised LDA (Downstream-sLDA)



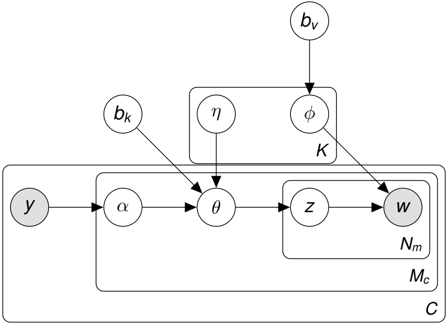
- Supervised LDA (sLDA) [2]
- $\overline{z}_{mk}$ is the average proportion of topic k in document m
- $y_m \sim \mathcal{N}(\eta_b + \eta^T \overline{z}_m, \sigma_y^2)$

Collectively Supervised LDA (Downstream-collective)



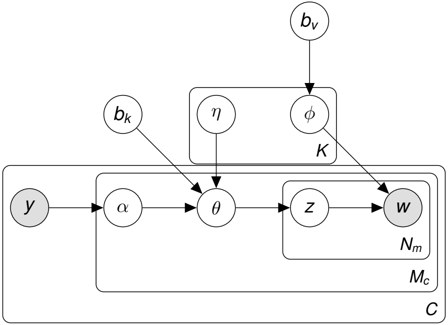
- Let $\overline{z}_{jk}$ be the average proportion of topic k in collection j
- $y_j \sim \mathcal{N}(\eta_b + \eta^T \overline{z}_j, \sigma_y^2)$
- Supervised LDA is a special case of this, where each document has its own unique collection ID

Dirichlet Multinomial Regression (Upstream)



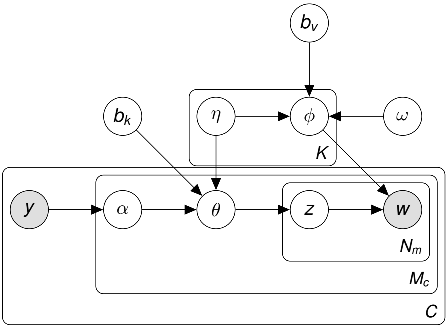
- Dirichlet-multinomial regression (DMR) [3]
- $\alpha_m = y_{c_m}$, feature value associated with document's collection c_m
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k); \theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
- $\tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$

DMR with adaptive supervision (Upstream-ada)



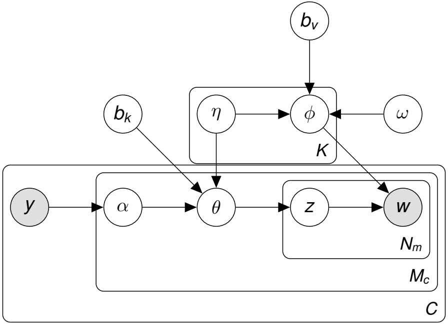
- $\alpha_m \sim \mathcal{N}(y_{c_m}, \sigma_\alpha^2)$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v)$; $\phi_k \sim \mathrm{Dirichlet}(\tilde{\phi}_k)$
- Document value can deviate from given input – can help infer likely values when supervision is noisy or missing.

DMR with word priors (Upstream-words)



- $\alpha_m = y_{c_m}$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v + \omega_v \eta_k)$
- Supervision affects priors over words. Extension to DMR known as SPRITE [7].

DMR + adaptive + word prior (Upstream-ada-words)



- Combined upstream model
- $\alpha_m \sim \mathcal{N}(y_{c_m}, \sigma_\alpha)$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v + \omega_v \eta_k)$

Surveys

BRFSS™

- Behavioral Risk Factor Surveillance System: annual survey by US federal government to learn about health/behavior of population.
- We selected three questions from BRFSS phone surveys:
  - **Guns**: Do you have a firearm in your house? (2001)
  - **Vaccines**: Have you had a flu shot in the past year? (2013)
  - **Smoking**: Are you a current smoker? (2013)
- Survey responses are aggregated at the level of US state.

Twitter Data

Dataset	Vocab	BRFSS
Guns	12,358	Owns firearm
Vaccines	13,451	Had flu shot
Smoking	13,394	Current smoker

- 100,000 tweets per dataset (filtered by relevant keywords)
  - collected between Dec. 2012 - Jan. 2015
- Identified as English using langid
  `https://github.com/saffsd/langid.py`
- Stopwords removed and low-frequency tokens excluded
- Location inferred using Carmen
  `https://github.com/mdredze/carmen-python`

Supervision

For each dataset:

- Each collection is defined as the set of tweets per US state
  - 50 collections
- Each collection's y_c value is the proportion respondents answering "Yes" to the BRFSS question

Predicting survey values:

- L2-regularized linear regression model
- Features: mean topic distributions θ per collection

Experiment Details

- Lots of hyperparameters – selected hyperparameters that maximized perplexity on heldout sample
- Optimized each model using Spearmint: `https://github.com/JasperSnoek/spearmint`
- Fit models using Gibbs sampling with AdaGrad for parameter (η) optimization
- Prediction task tuned with 5-fold cross validation: 80% train, 10% dev, 10% test.

Results

Features	Model	Guns		Vaccines		Smoking	
		RMSE	Perplexity	RMSE	Perplexity	RMSE	Perplexity
None	LDA	17.44	2313 (±52)	8.67	2524 (±20)	4.50	2118 (±5)
Survey	Upstream	15.37	1529 (±12)	6.54	1552 (±11)	3.41	1375 (±6)
	Upstream-words	11.50	**1429** (±22)	6.37	1511 (±57)	3.41	1374 (±2)
	Upstream-ada	11.48	1506 (±67)	5.82	**1493** (±49)	3.41	1348 (±6)
	Upstream-ada-words	**11.47**	1535 (±28)	7.20	1577 (±15)	**3.40**	1375 (±3)
	Downstream-sLDA	11.52	1561 (±22)	11.22	1684 (±7)	3.95	1412 (±3)
	Downstream-collective	12.81	1573 (±20)	9.17	1684 (±6)	4.35	1412 (±4)

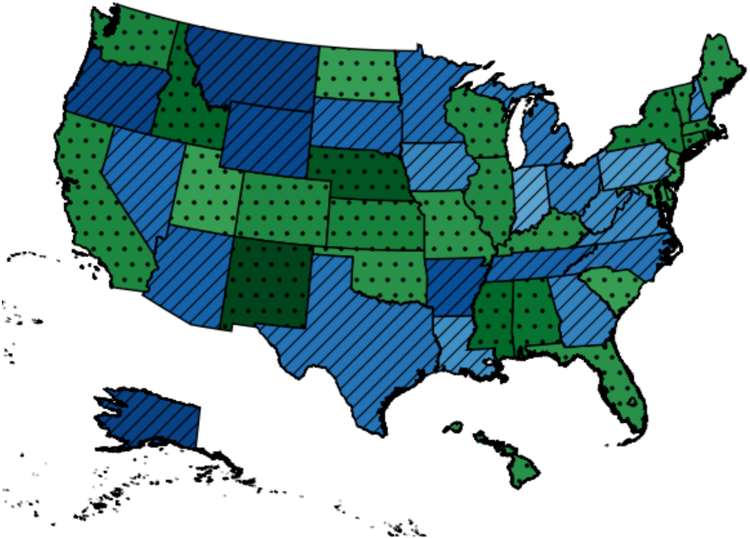
Use Case – Support for Universal Background Checks

- UBCs were a big US political issue in 2013, when national gun control legislation was floated
- We collected surveys on support for UBCs for 22 states from various polls (mostly Public Policy Polling)
- Baseline: use older 2001 survey of proportion households containing a firearm

Use Case – Support for Universal Background Checks

Features	Model	RMSE (2001 Y included)	RMSE (2001 Y omitted)
None	No model	7.26	7.59
	Bag of words	5.16	7.31
	LDA	6.40	7.59
Survey	Upstream-ada-words	**5.11**	**5.48**

Use Case – Support for Universal Background Checks



Code/Data

- Code and Data:
  `https://bitbucket.org/adrianbenton/sprite/`
- UBC Predictions:
  `https://github.com/abenton/collsuptmdata`

Questions?

References I

A. Culotta. Estimating county health statistics with Twitter. In *CHI*, 2014.

J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 121–128, 2008.

D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI*, 2008.

M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8), 2013.

B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

M. J. Paul and M. Dredze. You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*, pages 265–272, 2011.

M. J. Paul and M. Dredze. SPRITE: Generalizing topic models with structured priors. *Transactions of the Association for Computational Linguistics (TACL)*, 3:43–57, 2015.