Paul Felt, Eric Ringger, **Jordan Boyd-Graber**, and Kevin Seppi. **Making the Most of Crowdsourced Document Annotations: Confused Supervised LDA**. *Conference on Computational Natural Language Learning*, 2015, 10 pages.


```
    @inproceedings{Felt:Ringger:Boyd-Graber:Seppi-2015,
Title = {Making the Most of Crowdsourced Document Annotations: Confused Supervised {LDA}},
Author = {Paul Felt and Eric Ringger and Jordan Boyd-Graber and Kevin Seppi},
Booktitle = {Conference on Computational Natural Language Learning},
Year = {2015},
Location = {Beijing, China},
Url = {http://umiacs.umd.edu/~jbg//docs/2015_conll_cslda.pdf}
}
```


**This paper received the best paper award at CoNLL**

Links:

- Talk [http://techtalks.tv/talks/making-the-most-of-crowdsourced-document-annotations-confused-supervised-61895/]

# Making the Most of Crowdsourced Document Annotations: Confused Supervised LDA

**Paul Felt**
Dept. of Computer Science
Brigham Young University
paul.lewis.felt@gmail.com

**Eric K. Ringger**
Dept. of Computer Science
Brigham Young University
ringger@cs.byu.edu

**Jordan Boyd-Graber**
Dept. of Computer Science
University of Colorado Boulder
Jordan.Boyd.Graber@colorado.edu

**Kevin Seppi**
Dept. of Computer Science
Brigham Young University
kseppi@byu.edu

## Abstract

Corpus labeling projects frequently use low-cost workers from microtask marketplaces; however, these workers are often inexperienced or have misaligned incentives. Crowdsourcing models must be robust to the resulting systematic and non-systematic inaccuracies. We introduce a novel crowdsourcing model that adapts the discrete supervised topic model sLDA to handle multiple corrupt, usually conflicting (hence "confused") supervision signals. Our model achieves significant gains over previous work in the accuracy of deduced ground truth.

## 1 Modeling Annotators and Abilities

Supervised machine learning requires labeled training corpora, historically produced by laborious and costly annotation projects. Microtask markets such as Amazon's Mechanical Turk and Crowdflower have turned crowd labor into a commodity that can be purchased with relatively little overhead. However, crowdsourced judgments can suffer from high error rates. A common solution to this problem is to obtain multiple redundant human judgments, or annotations,[1] relying on the observation that, in aggregate, non-experts often rival or exceed experts by averaging over individual error patterns (Surowiecki, 2005; Snow et al., 2008; Jurgens, 2013).

A *crowdsourcing model* harnesses the wisdom of the crowd and infers labels based on the evidence of the available annotations, imperfect though they be. A common baseline crowdsourcing method aggregates annotations by *majority vote*, but this approach ignores important information. For example, some annotators are more reliable than others and their judgments ought to be upweighted accordingly. State-of-the-art crowdsourcing methods account for annotator expertise, often through a probabilistic formalism. Compounding the challenge, assessing unobserved annotator expertise is tangled with estimating ground truth from imperfect annotations; thus, joint inference of these interrelated quantities is necessary. State-of-the-art models also take the data into account, because data features can help ratify or veto human annotators.

We introduce a model that improves on state of the art crowdsourcing algorithms by modeling not only the annotations but also the features of the data (e.g., words in a document). Section 2 identifies modeling deficiencies affecting previous work and proposes a solution based on topic modeling; Section 2.4 presents inference for the new model. Experiments that contrast the proposed model with select previous work on several text classification datasets are presented in Section 3. In Section 4 we highlight additional related work.

## 2 Latent Representations that Reflect Labels and Confusion

Most crowdsourcing models extend the item-response model of Dawid and Skene (1979). The Bayesian version of this model, referred to here as ITEMRESP, is depicted in Figure 1. In the generative story for this model, a confusion matrix $\gamma_j$ is drawn for each human annotator $j$. Each row $\gamma_{jc}$ of the confusion matrix $\gamma_j$ is drawn from

[1] In this paper, we call human judgments *annotations* to distinguish them from gold standard class *labels*.

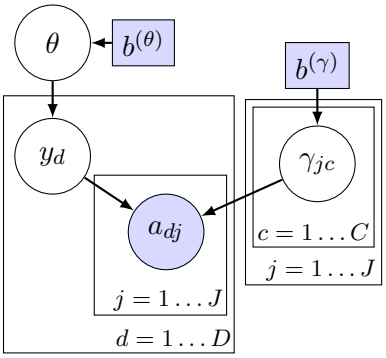



Figure 1: **ITEMRESP** as a plate diagram. Round nodes are random variables. Rectangular nodes are free parameters. Shaded nodes are observed. $D, J, C$ are the number of documents, annotators, and classes, respectively.

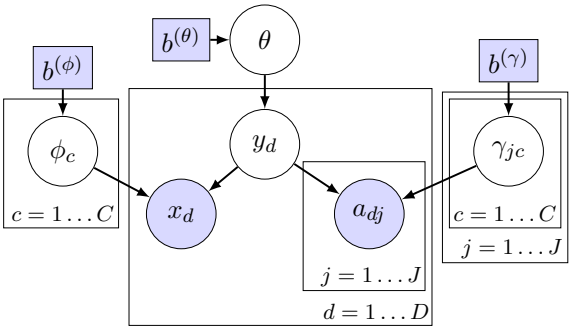



Figure 2: **MOMRESP** as a plate diagram. $|x_d| = V$, the size of the vocabulary. Documents with similar feature vectors $x$ tend to share a common label $y$. Reduces to mixture-of-multinomials clustering when no annotations $a$ are observed.

a symmetric Dirichlet distribution $Dir(b_{jc}^{(\gamma)})$ and encodes a categorical probability distribution over label classes that annotator $j$ is apt to choose when presented with a document whose true label is $c$. Then for each document $d$ an unobserved document label $y_d$ is drawn. Annotations are generated as annotator $j$ corrupts the true label $y_d$ according to the categorical distribution $Cat(\gamma_{jy_d})$.

### 2.1 Leveraging Data

Some extensions to ITEMRESP model the features of the data (e.g., words in a document). Many data-aware crowdsourcing models condition the labels on the data (Jin and Ghahramani, 2002; Raykar et al., 2010; Liu et al., 2012; Yan et al., 2014), possibly because discriminative classifiers dominate supervised machine learning. Others model the data generatively (Bragg et al., 2013; Lam and Stork, 2005; Felt et al., 2014; Simpson and Roberts, 2015). Felt et al. (2015) argue that generative models are better suited than conditional models to crowdsourcing scenarios because generative models often learn faster than their conditional counterparts (Ng and Jordan, 2001)—especially early in the learning curve. This advantage is amplified by the annotation noise typical of crowdsourcing scenarios.

Extensions to ITEMRESP that model document features generatively tend to share a common high-level architecture. After the document class label $y_d$ is drawn for each document $d$, features are drawn from class-conditional distributions. Felt et al. (2015) identify the MOMRESP model, reproduced in Figure 2, as a strong representative of generative crowdsourcing models. In MOMRESP, the feature vector $x_d$ for document $d$ is drawn from the multinomial distribution with parameter vector $\phi_{y_d}$. This class-conditional multinomial model of the data inherits many of the strengths and weaknesses of the naïve Bayes model that it resembles. Strengths include easy inference and a strong inductive bias which helps the model be robust to annotation noise and scarcity. Weaknesses include overly strict conditional independence assumptions among features, leading to overconfidence in the document model and thereby causing the model to overweight feature evidence and underweight annotation evidence. This imbalance can result in degraded performance in the presence of high quality or many annotations.

### 2.2 Confused Supervised LDA (CSLDA)

We solve the problem of imbalanced feature and annotation evidence observed in MOMRESP by replacing the class-conditional structure of previous generative crowdsourcing models with a richer generative story where documents are drawn first and class labels $y_d$ are obtained afterwards via a log-linear mapping. This move towards conditioning classes on documents content is sensible because in many situations document content is authored first, whereas label structure is not imposed until afterwards. It is plausible to assume that there will exist some mapping from a latent document structure to the desired document label distinctions. Moreover, by jointly modeling topics and the mapping to labels, we can learn the latent document representations that best explain how best to predict and correct annotator errors.

| Term | Definition |
|---|---|
| $N_d$ | Size of document $d$ |
| $N_{dt}$ | $\sum_n \mathbb{1}(z_{dn} = t)$ |
| $N_t$ | $\sum_{d,n} \mathbb{1}(z_{dn} = t)$ |
| $N_{jcc'}$ | $\sum_d a_{djc'} \mathbb{1}(y_d = c)$ |
| $N_{jc}$ | $\langle N_{jc1} \cdots N_{jcC} \rangle$ |
| $N_{vt}$ | $\sum_{dn} \mathbb{1}(w_{dn} = v \wedge z_{dn} = t)$ |
| $N_t$ | $\sum_{dn} \mathbb{1}(z_{dn} = t)$ |
| $\hat{N}$ | Count excludes variable being sampled |
| $\bar{z}_d$ | Vector where $\bar{z}_{dt} = \frac{1}{N_d} \sum_n \mathbb{1}(z_{dn} = t)$ |
| $\hat{\bar{z}}_d$ | Excludes the $z_{dn}$ being sampled |

Table 1: Definition of counts and select notation. $\mathbb{1}(\cdot)$ is the indicator function.

We call our model **confused supervised LDA** (**csLDA**, Figure 3), based on supervised topic modeling. Latent Dirichlet Allocation (Blei et al., 2003, LDA) models text documents as admixtures of word distributions, or topics. Although pre-calculated LDA topics as features can inform a crowdsourcing model (Levenberg et al., 2014), supervised LDA (sLDA) provides a principled way of incorporating document class labels and topics into a single model, allowing topic variables and response variables to co-inform one another in joint inference. For example, when sLDA is given movie reviews labeled with sentiment, inferred topics cluster around sentiment-heavy words (Mcauliffe and Blei, 2007), which may be quite different from the topics inferred by unsupervised LDA. One way to view csLDA is as a discrete sLDA in settings with noisy supervision from multiple, imprecise annotators.

The generative story for csLDA is:

1. Draw per-topic word distributions $\phi_t$ from $Dir(b^{(\theta)})$.
2. Draw per-class regression parameters $\eta_c$ from $Gauss(\mu, \Sigma)$.
3. Draw per-annotator confusion matrices $\gamma_j$ with row $\gamma_{jc}$ drawn from $Dir(b^{(\gamma)}_{jc})$.
4. For each document $d$,
   (a) Draw topic vector $\theta_d$ from $Dir(b^{(\theta)})$.
   (b) For each token position $n$, draw topic $z_{dn}$ from $Cat(\theta_d)$ and word $w_{dn}$ from $Cat(\phi_{z_{dn}})$.
   (c) Draw class label $y_d$ with probability proportional to $\exp[\eta_{y_d}^\intercal \bar{z}_d]$.
   (d) For each annotator $j$ draw annotation vector $a_{dj}$ from $\gamma_{jy_d}$.

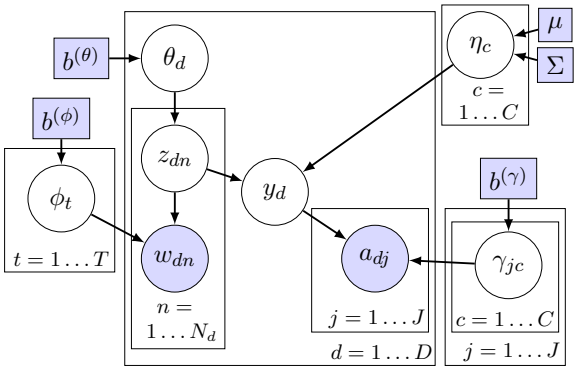



Figure 3: **csLDA** as a plate diagram. $D, J, C, T$ are the number of documents, annotators, classes, and topics, respectively. $N_d$ is the size of document $d$. $|\phi_t| = V$, the size of the vocabulary. $\eta_c$ is a vector of regression parameters. Reduces to LDA when no annotations $a$ are observed.

### 2.3 Stochastic EM

We use stochastic expectation maximization (EM) for posterior inference in csLDA, alternating between sampling values for topics $z$ and document class labels $y$ (the E-step) and optimizing values of regression parameters $\eta$ (the M-step). To sample $z$ and $y$ efficiently, we derive the full conditional distributions of $z$ and $y$ in a collapsed model where $\theta$, $\phi$, and $\gamma$ have been analytically integrated out. Omitting multiplicative constants, the collapsed model joint probability is

$$p(z, w, y, a|\eta) = p(z)p(w|z)p(y|z, \eta)p(a|y) \quad (1)$$
$$\propto M(a) \cdot \left( \prod_d B(N_d + b^{(\theta)}) \right) \cdot \left( \prod_t B(N_t + b^{(\phi)}_t) \right)$$
$$\cdot \left( \prod_d \frac{\exp(\eta_{y_d}^\intercal \bar{z}_d)}{\sum_c \exp(\eta_c^\intercal \bar{z}_d)} \right) \cdot \left( \prod_j \prod_c B(N_{jc} + b^{(\gamma)}_{jc}) \right)$$

where $B(\cdot)$ is the Beta function (multivariate as necessary), counts $N$ and related symbols are defined in Table 1, and $M(a) = \prod_{d,j} M(a_{dj})$ where $M(a_{dj})$ is the multinomial coefficient.

Simplifying Equation 1 yields full conditionals for each word $z_{dn}$,

$$p(z_{dn} = t|\hat{z}, w, y, a, \eta) \propto \left( \hat{N}_{dt} + b^{(\theta)}_t \right) \quad (2)$$
$$\cdot \frac{\hat{N}_{w_{dn}t} + b^{(\phi)}_{w_{dn}}}{\hat{N}_t + |b^{(\phi)}|_1} \cdot \frac{\exp(\frac{\eta_{y_d t}}{N_d})}{\sum_c \exp(\frac{\eta_{ct}}{N_d} + \eta_c^\intercal \hat{\bar{z}}_d)},$$

and similarly for document label $y_d$:

$$p(y_d = c|z, w, y, a, \eta) \propto \frac{\exp(\eta_c^\intercal \bar{z}_d)}{\sum_{c'} \exp(\eta_{c'}^\intercal \bar{z}_d)} \cdot \prod_j \frac{\prod_{c'} \left(\hat{N}_{jcc'} + b^{(\gamma)}\right)^{\overline{a_{djc'}}}}{\left(\sum_{c'} \hat{N}_{jcc'} + b^{(\gamma)}_{jcc'}\right)^{\overline{\sum_{c'} a_{djc'}}}}, \quad (3)$$

where $x^{\overline{k}} \triangleq x(x+1)\cdots(x+k-1)$ is the rising factorial. In Equation 2 the first and third terms are independent of word $n$ and can be cached at the document level for efficiency.

For the M-step, we train the regression parameters $\eta$ (containing one vector per class) by optimizing the same objective function as for training a logistic regression classifier, assuming that class $y$ is given:

$$p(y = c|z, \eta) = \prod_d \frac{\exp(\eta_c^\intercal \bar{z}_d)}{\sum_{c'} \exp(\eta_{c'}^\intercal \bar{z}_d)}. \quad (4)$$

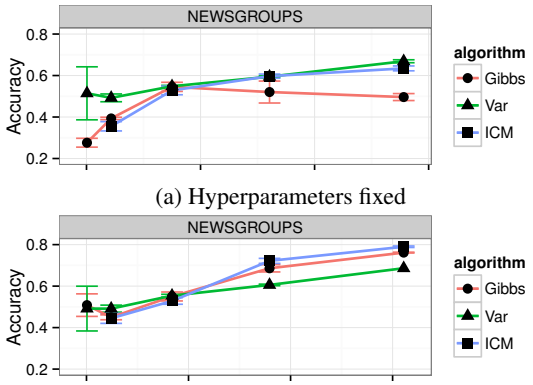
We optimize the objective (Equation 4) using L-BFGS and a regularizing Gaussian prior with $\mu = 0, \sigma^2 = 1$.

While EM is sensitive to initialization, CSLDA is straightforward to initialize. Majority vote is used to set initial $y$ values $\tilde{y}$. Corresponding initial values for $z$ and $\eta$ are obtained by clamping $y$ to $\tilde{y}$ and running stochastic EM on $z$ and $\eta$.

### 2.4 Hyperparameter Optimization

Ideally, we would test CSLDA performance under all of the many algorithms available for inference in such a model. Although that is not feasible, Asuncion et al. (2009) demonstrate that hyperaparameter optimization in LDA topic models helps to bring the performance of alternative inference algorithms into approximate agreement. Accordingly, in Section 2.4 we implement hyperparameter optimization for CSLDA to make our results as general as possible.

Before moving on, however, we take a moment to validate that the observation of Asuncion et al. generalizes from LDA to the ITEMRESP model, which, together with LDA, comprises CSLDA. Figure 4 demonstrates that three ITEMRESP inference algorithms, Gibbs sampling (Gibbs), mean-field variational inference (Var), and the iterated conditional modes algorithm (ICM) (Besag, 1986), are brought into better agreement after optimizing their hyperparameters via grid search. That




(a) Hyperparameters fixed




(b) Hyperparameters optimized via grid search on validation data

Figure 4: Differences among the inferred label accuracy learning curves of three ITEMRESP inference algorithms are reduced when hyperparameters are optimized.

is, the algorithms in Figure 4b are in better agreement, particularly near the extremes, than the algorithms in Figure 4a. This difference is subtle, but it holds to an equal and greater extent in other simulation conditions we tested (experiment details are similar to those reported in Section 3).

**Fixed-point Hyperparameter Updates**

Although a grid search is effective, it is not practical for a model with many hyperparameters such as CSLDA. For efficiency, therefore, we use the fixed-point updates of Minka (2000). Our updates differ slightly from Minka's since we tie hyperparameters, allowing them to be learned more quickly from less data. In our implementation the matrices of hyperparameters $b^{(\phi)}$ and $b^{(\theta)}$ over the Dirichlet-multinomial distributions are completely tied such that $b^{(\phi)}_{tv} = b^{(\phi)} \forall t, v$ and $b^{(\theta)}_t = b^{(\theta)} \forall t$. This leads to

$$b^{(\phi)} \leftarrow b^{(\phi)} \cdot \frac{\sum_{t,v}[\Psi(N_{tv} + b^{(\phi)})] - TV\Psi(b^{(\phi)})}{V[\Psi(N_t + Vb^{(\phi)}) - \Psi(Vb^{(\phi)})]} \quad (5)$$

and

$$b^{(\theta)} \leftarrow b^{(\theta)} \cdot \frac{\sum_{d,t}[\Psi(N_{dt} + b^{(\theta)})] - NT\Psi(b^{(\theta)})}{T[\Psi(N_d + Tb^{(\theta)}) - \Psi(Tb^{(\theta)})]}. \quad (6)$$

The updates for $b^{(\gamma)}$ are slightly more involved since we choose to tie the diagonal entries $b^{(\gamma)}_d$ and separately the off-diagonal entries $b^{(\gamma)}_o$, updating each separately:

$$b^{(\gamma)}_d \leftarrow b^{(\gamma)}_d \cdot \frac{\sum_{j,c}[\Psi(N_{jcc} + b^{(\gamma)}_d)] - JC\Psi(b^{(\gamma)}_d)}{Z(b^{(\gamma)})} \quad (7)$$

and $b_o^{(\gamma)} \leftarrow$

$$b_o^{(\gamma)} \cdot \frac{\sum_{j,c,c' \neq c} [\Psi(N_{jcc'} + b_o^{(\gamma)})] - JC(C-1)\Psi(b_o^{(\gamma)})]}{(C-1)Z(b^{(\gamma)})} \quad (8)$$

where

$$Z(b^{(\gamma)}) = \sum_{j,c} [\Psi(N_{jc} + b_d^{(\gamma)} + (C-1)b_o^{(\gamma)})] - JC\Psi(b_d^{(\gamma)} + (C-1)b_o^{(\gamma)}).$$

As in the work of Asuncion et al. (2009), we add an algorithmic gamma prior ($b^{(\cdot)} \sim G(\alpha, \beta)$) for smoothing by adding $\frac{\alpha-1}{b^{(\cdot)}}$ to the numerator and $\beta$ to the denominator of Equations 5-8. Note that these algorithmic gamma "priors" should not be understood as first-class members of the CSLDA model (Figure 3). Rather, they are regularization terms that keep our hyperparameter search algorithm from straying towards problematic values such as 0 or $\infty$.

## 3 Experiments

For all experiments we set CSLDA's number of topics $T$ to 1.5 times the number of classes in each dataset. We found that model performance was reasonably robust to this parameter. Only when $T$ drops below the number of label classes does performance suffer. As per Section 2.3, $z$ and $\eta$ values are initialized with 500 rounds of stochastic EM, after which the full model is updated with 1000 additional rounds. Predictions are generated by aggregating samples from the last 100 rounds (the mode of the approximate marginal posterior).

We compare CSLDA with (1) a majority vote baseline, (2) the ITEMRESP model, and representatives of the two main classes of data-aware crowdsourcing models, namely (3) data-generative and (4) data-conditional. MOMRESP represents a typical data-generative model (Bragg et al., 2013; Felt et al., 2014; Lam and Stork, 2005; Simpson and Roberts, 2015). Data-conditional approaches typically model data features conditionally using a log-linear model (Jin and Ghahramani, 2002; Raykar et al., 2010; Liu et al., 2012; Yan et al., 2014). For the purposes of this paper, we refer to this model as LOGRESP. For ITEMRESP, MOMRESP, and LOGRESP we use the variational inference methods presented by Felt et al. (2015). Unlike that paper, in this work we have augmented inference with the in-line hyperparameter updates described in Section 2.4.

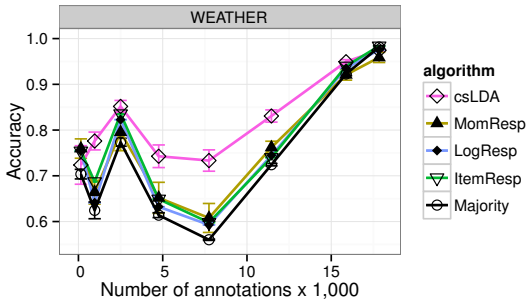



Figure 5: Inferred label accuracy of models on sentiment-annotated weather tweets.

### 3.1 Human-generated Annotations

To gauge the effectiveness of data-aware crowdsourcing models, we use the sentiment-annotated tweet dataset distributed by CrowdFlower as a part of its "data for everyone" initiative.[2] In the "Weather Sentiment" task, 20 annotators judged the sentiment of 1000 tweets as either positive, negative, neutral, or unrelated to the weather. In the secondary "Weather Sentiment Evaluated" task, 10 additional annotators judged the correctness of each consensus label. We construct a gold standard from the consensus labels that were judged to be correct by 9 of the 10 annotators in the secondary task.

Figure 5 plots learning curves of the accuracy of model-inferred labels as annotations are added (ordered by timestamp). All methods, including majority vote, converge to roughly the same accuracy when all 20,000 annotations are added. When fewer annotations are available, statistical models beat majority vote, and CSLDA is considerably more accurate than other approaches. Learning curves are bumpy because annotation order is not random and because inferred label accuracy is calculated only over documents with at least one annotation. Learning curves collectively increase when average annotation depth (the number of annotations per item) increases and decrease when new documents are annotated and average annotation depth decreases. CSLDA stands out by being more robust to these changes than other algorithms, and also by maintaining a higher level of accuracy across the board. This is important because high accuracy using fewer annotations translates to decreased annotations costs.

[2] http://www.crowdflower.com/data-for-everyone

| | $D$ | $C$ | $V$ | $\frac{1}{N}\sum_d N_d$ |
|---|---|---|---|---|
| 20 News | 16,995 | 20 | 22,851 | 111 |
| WebKB | 3,543 | 4 | 5,781 | 131 |
| Reuters8 | 6,523 | 8 | 6,776 | 53 |
| Reuters52 | 7,735 | 52 | 5,579 | 58 |
| CADE12 | 34,801 | 12 | 41,628 | 110 |
| Enron | 3,854 | 32 | 14,069 | 431 |

Table 2: Dataset statistics. $D$ is number of documents, $C$ is number of classes, $V$ is number of features, and $\frac{1}{N}\sum_d N_d$ is average document size. Values are calculated after setting aside 15% as validation data and doing feature selection.

### 3.2 Synthetic Annotations

Datasets including both annotations and gold standard labels are in short supply. Although plenty of text categorization datasets have been annotated, common practice reflects that initial noisy annotations be discarded and only consensus labels be published. Consequently, we follow previous work in achieving broad validation by constructing synthetic annotators that corrupt known gold standard labels. We base our experimental setup on the annotations gathered by Felt et al. (2015),[3] who paid CrowdFlower annotators to relabel 1000 documents from the well-known 20 Newsgroups classification dataset. In that experiment, 136 annotators contributed, each instance was labeled an average of 6.9 times, and annotator accuracies were distributed approximately according to a $Beta(3.6, 5.1)$ distribution. Accordingly we construct 100 synthetic annotators, each parametrized by an accuracy drawn from $Beta(3.6, 5.1)$ and with errors drawn from a symmetric Dirichlet $Dir(1)$. Datasets are annotated by selecting an instance (at random without replacement) and then selecting $K$ annotators (at random without replacement) to annotate it before moving on. We choose $K = 7$ to mirror the empirical average in the CrowdFlower annotation set.

We evaluate on six text classification datasets, summarized in Table 2. The 20 Newsgroups, WebKB, Cade12, Reuters8, and Reuters52 datasets are described in more detail by Cardoso-Cachopo (2007). The LDC-labeled Enron emails dataset is described by Berry et al. (2001). Each dataset is

[3]The dataset is available via git at `git://nlp.cs.byu.edu/plf1/crowdflower-newsgroups.git`

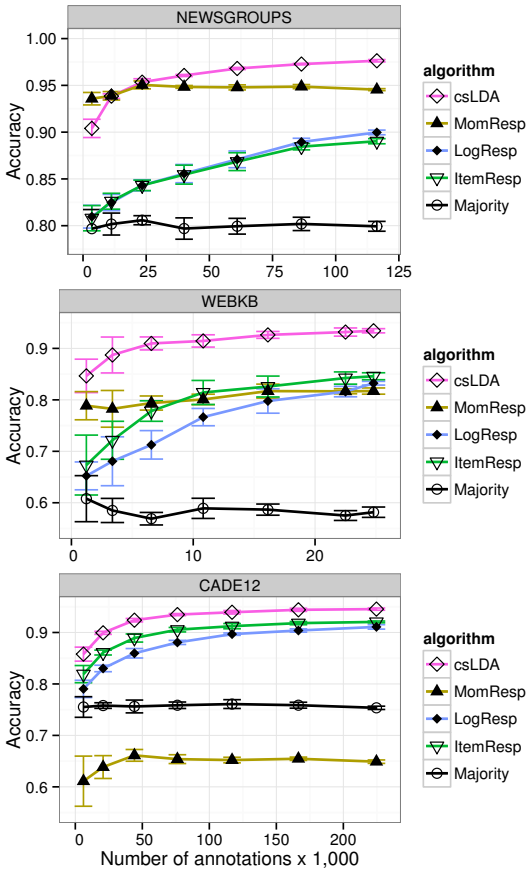



Figure 6: Inferred label accuracy of models on synthetic annotations. The first instance is annotated 7 times, then the second, and so on.

preprocessed via Porter stemming and by removal of the stopwords from MALLET's stopword list (McCallum, 2002). Features occurring fewer than 5 times in the corpus are discarded. In the case of MOMRESP, features are fractionally scaled so that each document is the same length, in keeping with previous work in multinomial document models (Nigam et al., 2006).

Figure 6 plots learning curves on three representative datasets (Enron resembles Cade12, and the Reuters datasets resemble WebKB). CSLDA consistently outperforms LOGRESP, ITEMRESP, and majority vote. The generative models (CSLDA and MOMRESP) tend to excel in low-annotation portions of the learning curve, partially because generative models tend to converge quickly and partially because generative models naturally learn from unlabeled documents (i.e., semi-supervision). However, MOMRESP tends to quickly reach a performance plateau after which additional annotations do little good. The performance of MOMRESP is also highly dataset de-

| 95% Accuracy | csLDA | MomResp | LogResp | ItemResp | Majority |
|---|---|---|---|---|---|
| 20 News | **85 (5.0x)** | 150 (8.8x) | 152 (8.9x) | 168 (9.9x) | 233 (13.7x) |
| WebKB | **31 (8.8x)** | - | 46 (13.0x) | 46 (13.0x) | - |
| Reuters8 | **25 (3.8x)** | - | 73 (11.2x) | 62 (9.5x) | - |
| Reuters52 | **33 (4.3x)** | 73 (9.4x) | 67.5 (8.7x) | 60 (7.8x) | 87 (11.2x) |
| CADE12 | **250 (7.2x)** | - | 295 (8.5x) | 290 (8.3x) | 570 (16.4x) |
| Enron | **31 (8.0x)** | - | 40 (10.4x) | 38 (9.9x) | 47 (12.2x) |

Table 3: The number of annotations $\times 1000$ at which the algorithm reaches 95% inferred label accuracy on the indicated dataset (average annotations per instance are in parenthesis). All instances are annotated once, then twice, and so on. Empty entries ('-') do not reach 95% even with 20 annotations per instance.

pendent: it is good on 20 Newsgroups, mediocre on WebKB, and poor on CADE12. By contrast, csLDA is relatively stable across datasets.

To understand the different behavior of the two generative models, recall that MomResp is identical to ItemResp save for its multinomial data model. Indeed, the equations governing inference of label $y$ in MomResp simply sum together terms from an ItemResp model and terms from a mixture of multinomials clustering model (and for reasons explained in Section 2.1, the multinomial data model terms tend to dominate). Therefore when MomResp diverges from ItemResp it is because MomResp is attracted toward a $y$ assignment that satisfies the multinomial data model, grouping similar documents together. This can both help and hurt. When data clusters and label classes are misaligned, MomResp falters (as in the case of the Cade12 dataset). In contrast, csLDA's flexible mapping from topics to labels is less sensitive: topics can diverge from label classes so long as there exists some linear transformation from the topics to the labels.

Many corpus annotation projects are not complete until the corpus achieves some target level of quality. We repeat the experiment reported in Figure 6, but rather than simulating seven annotations for each instance before moving on, we simulate one annotation for each instance, then two, and so on until each instance in the dataset is annotated 20 times. Table 3 reports the minimal number of annotations before an algorithm's inferred labels reach an accuracy of 95%, a lofty goal that can require significant amounts of annotation when using poor quality annotations. csLDA achieves 95% accuracy with fewer annotations, corresponding to reduced annotation cost.

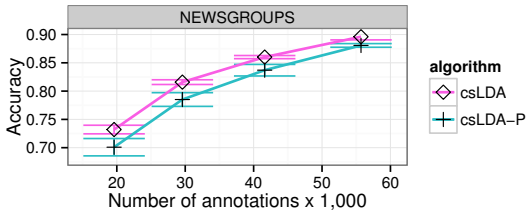


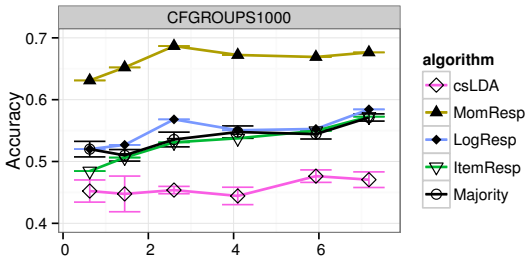

Figure 7: Joint inference for csLDA vs pipeline inference (csLDA-P).

### 3.3 Joint vs Pipeline Inference

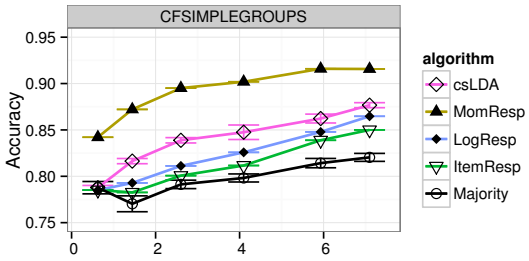
To isolate the effectiveness of joint inference in csLDA, we compare against the pipeline alternative where topics are inferred first and then held constant during inference (Levenberg et al., 2014). Joint inference yields modest but consistent benefits over a pipeline approach. Figure 7 highlights a portion of the learning curve on the Newsgroups dataset (based on the experiments summarized in Table 3). This trend holds across all of the datasets that we examined.

### 3.4 Error Analysis

Class-conditional models like MomResp include a feature that data-conditional models like csLDA lack: an explicit prior over class prevalence. Figure 8a shows that csLDA performs poorly on the CrowdFlower-annotated Newsgroups documents described at the beginning of Section 3 (not the synthetic annotations). Error analysis uncovers that csLDA lumps related classes together in this dataset. This is because annotators could specify up to 3 simultaneous labels for each annotation, so that similar labels (e.g., "talk.politics.misc" and "talk.politics.mideast") are usually chosen in blocks. Suppose each member of a set of documents with similar topical content is annotated




(a) Original data




(b) After combining frequently co-annotated label classes

Figure 8: An illustrative failure case. CSLDA, lacking a class label prior, prefers to combine label classes that are highly co-annotated.

with both label A and B. In this scenario it is apparent that CSLDA will achieve its best fit by inferring all documents to have the same label either A or B. By contrast, MOMRESP's uniform prior distribution over $\theta$ leads it to prefer solutions with a balance of A and B.

The hypothesis that class combination explains CSLDA's performance is supported by Figure 8b, which shows that CSLDA recovers after combining the classes that were most frequently co-annotated. We greedily combine label class pairs to maximize Krippendorf's $\alpha$ until only 10 labels were left: "alt.atheism," religion, and politics classes were combined; also, "sci.electronics" and the computing classes. The remaining eight classes were unaltered. However, one could also argue that the original behavior of CSLDA is in some ways desirable. That is, if two classes of documents are mostly the same both topically and in terms of annotator decisions, perhaps those classes ought to be collapsed. We are not overly concerned that MOMRESP beats CSLDA in Figure 8, since this result is consistent with early relative performance in simulation.

## 4 Additional Related Work

This section reviews related work not already discussed. A growing body of work extends the item-response model to account for variables such as item difficulty (Whitehill et al., 2009; Passonneau and Carpenter, 2013; Zhou et al., 2012), annotator trustworthiness (Hovy et al., 2013), correlation among various combinations of these variables (Zhou et al., 2014), and change in annotator behavior over time (Simpson and Roberts, 2015).

Welinder et al. (2010) carefully model the process of annotating objects in images, including variables for item difficulty, item class, and class-conditional perception noise. In follow-up work, Liu et al. (2012) demonstrate that similar levels of performance can be achieved with the simple item-response model by using variational inference rather than EM. Alternative inference algorithms have been proposed for crowdsourcing models (Dalvi et al., 2013; Ghosh et al., 2011; Karger et al., 2013; Zhang et al., 2014). Some crowdsourcing work regards labeled data not as an end in itself, but rather as a means to train classifiers (Lin et al., 2014). The fact-finding literature assigns trust scores to assertions made by untrusted sources (Pasternack and Roth, 2010).

## 5 Conclusion and Future Work

We describe CSLDA, a generative, data-aware crowdsourcing model that addresses important modeling deficiencies identified in previous work. In particular, CSLDA handles data in which the natural document clusters are at odds with the intended document labels. It also transitions smoothly from situations in which few annotations are available to those in which many annotations are available. Because of the flexible mapping in CSLDA to class labels, many structural variants are possible in future work. For example, this mapping could depend not just on inferred topical content but also directly on data features (c.f. Nguyen et al. (2013)) or learned embedded feature representations.

The large number of parameters in the learned confusion matrices of crowdsourcing models present difficulty at scale. This could be addressed by modeling structure both inside of the annotators and classes. Redundant annotations give unique insights into both inter-annotator and inter-class relationships and could be used to induce annotator or label class hierarchies with parsimonious representations. Simpson et al. (2013) identify annotator clusters using community detection algorithms but do not address annotator hierarchy or scalable confusion representations.

**Acknowledgments** This work was supported by the collaborative NSF Grant IIS-1409739 (BYU) and IIS-1409287 (UMD). Boyd-Graber is also supported by NSF grants IIS-1320538 and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

## References

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. 2009. On smoothing and inference for topic models. In *Proceedings of Uncertainty in Artificial Intelligence*.

M. W. Berry, M. Browne, and B. Signer. 2001. Topic annotated Enron email data set. *Linguistic Data Consortium*.

J. Besag. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

J. Bragg, Mausam, and D. Weld. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI Conference on Human Computation and Crowdsourcing*.

A. Cardoso-Cachopo. 2007. *Improving Methods for Single-label Text Categorization*. Ph.D. thesis, Universidade Tecnica de Lisboa.

N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. 2013. Aggregating crowdsourced binary ratings. In *Proceedings of World Wide Web Conference*.

A.P. Dawid and A.M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28.

P. Felt, R. Haertel, E. Ringger, and K. Seppi. 2014. MomResp: A Bayesian model for multi-annotator document labeling. In *International Language Resources and Evaluation*.

P. Felt, E. Ringger, K. Seppi, and R. Haertel. 2015. Early gains matter: A case for preferring generative over discriminative crowdsourcing models. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

A. Ghosh, S. Kale, and P. McAfee. 2011. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*.

D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy. 2013. Learning whom to trust with MACE. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

R. Jin and Z. Ghahramani. 2002. Learning with multiple labels. In *Proceedings of Advances in Neural Information Processing Systems*, pages 897–904.

D. Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of NAACL-HLT*, pages 556–562.

D. Karger, S. Oh, and D. Shah. 2013. Efficient crowdsourcing for multi-class labeling. In *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 81–92. ACM.

C. P. Lam and D. G. Stork. 2005. Toward optimal labeling strategy under multiple unreliable labelers. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*.

A. Levenberg, S. Pulman, K. Moilanen, E. Simpson, and S. Roberts. 2014. Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering*, 3(2):109–115.

C. Lin, Mausam, and D. Weld. 2014. To re (label), or not to re (label). In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Q. Liu, J. Peng, and A. Ihler. 2012. Variational inference for crowdsourcing. In *Proceedings of Advances in Neural Information Processing Systems*.

J. Mcauliffe and D. Blei. 2007. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

A. McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

T. Minka. 2000. Estimating a Dirichlet distribution.

A. Ng and M. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. *Proceedings of Advances in Neural Information Processing Systems*.

V. Nguyen, J. Boyd-Graber, and P. Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*.

K. Nigam, A. McCallum, and T. Mitchell. 2006. Semi-supervised text classification using EM. *Semi-Supervised Learning*, pages 33–56.

R. Passonneau and B. Carpenter. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195.

J. Pasternack and D. Roth. 2010. Knowing what to believe (when you already know something). In *Proceedings of International Conference on Computational Linguistics*.

V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Edwin Simpson and Stephen Roberts. 2015. Bayesian methods for intelligent task assignment in crowdsourcing systems. In *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*, pages 1–32. Springer.

E. Simpson, S. Roberts, I. Psorakis, and A. Smith. 2013. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*. ACL.

J. Surowiecki. 2005. *The Wisdom of Crowds*. Random House LLC.

P. Welinder, S. Branson, P. Perona, and S. Belongie. 2010. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *NIPS*, 22:2035–2043.

Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy. 2014. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327.

Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. 2014. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems 27*, pages 1260–1268. Curran Associates, Inc.

D. Zhou, J. Platt, S. Basu, and Y. Mao. 2012. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, volume 25, pages 2204–2212.

D. Zhou, Q. Liu, J. Platt, and C. Meek. 2014. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of the International Conference of Machine Learning*.