Evaluating Supervised Topic Models in the Presence of OCR Errors

Daniel Walker, Eric Ringger, and Kevin Seppi

Natural Language Processing Lab, Computer Science Dept.
Brigham Young University, Provo, UT, USA
{danl4,ringger,k}@cs.byu.edu

ABSTRACT

Supervised topic models are promising tools for text analytics that simultaneously model topical patterns in document collections and relationships between those topics and document metadata, such as timestamps. We examine empirically the effect of OCR noise on the ability of supervised topic models to produce high quality output through a series of experiments in which we evaluate three supervised topic models and a naive baseline on synthetic OCR data having various levels of degradation and on real OCR data from two different decades. The evaluation includes experiments with and without feature selection. Our results suggest that supervised topic models are no better, or at least not much better in terms of their robustness to OCR errors, than unsupervised topic models and that feature selection has the mixed result of improving topic quality while harming metadata prediction quality. For users of topic modeling methods on OCR data, supervised topic models do not yet solve the problem of finding better topics than the original unsupervised topic models.

1. INTRODUCTION

As text data becomes available in massive quantities, it becomes increasingly difficult for human curators to manually catalog and index modern document collections. Topic models, such as LDA,[1] have emerged as one notable method for automatically discovering the topics discussed in a given document corpus. These models typically contain a latent topic label for each token, a latent distribution over topics for each document, and a latent distribution over vocabulary for each topic. Given a collection of documents, tools from Bayesian statistics (such as Gibbs sampling and variational inference) are used to infer values of these latent variables, the topic labels along with the topics themselves, in an unsupervised fashion. The topics thus discovered using topic models have the potential to facilitate browsing and the discovery of topical patterns and trends. This type of analysis has grown in popularity recently as inference on models containing large numbers of latent variables has become feasible due to algorithmic and computational advances.

Building on topic models, another class of document models called supervised topic models discovers topical labelings of words but also jointly models continuous metadata associated with the documents. For example, many documents have a known creation date. Also, product reviews often include a numerical rating summarizing the sentiment of the review's author. These models are especially promising for scholars of historical document collections because they allow for the discovery of patterns in the correlations between topics and metadata. For example, they can be used to trace the evolution of topics over time and to explore what the mention of a particular topic means in terms of the sentiment being expressed by an author. Many examples of supervised topic models exist in the literature: sLDA,[2] Topics Over Time,[3] Dirichlet Multinomial Regression,[4] and Topics Over Nonparametric Time.[5]

These models have been typically trained and evaluated on relatively clean datasets, but the modern explosion of text data includes vast amounts of historical documents made available by means of Optical Character Recognition (OCR), which can introduce significant numbers of errors. Undertakings to produce such data include the Google Books, Internet Archive, and HathiTrust projects. Due to their nature, these collections often lack topical annotations and may contain documents with missing or incorrect metadata as evidenced by the many publicized problems with the metadata associated with Google Books documents.[*] Depending on the age of a document and the way in which it was created, the OCR process results in text containing many types of noise, including character-level errors, which distort the counts of words and co-occurrences of words. However, finding good estimates for the parameters of supervised topic models requires that these counts be accurate. It is ostensible, therefore, that model quality must suffer, especially since the performance of completely unsupervised topic models are known to degrade in the presence of OCR errors.[6]

[*] http://chronicle.com/article/Googles-Book-Search-A/48245/

Document Recognition and Retrieval XX, edited by Richard Zanibbi, Bertrand Coüasnon,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8658, 865812 · © 2013 SPIE-IS&T
CCC code: 0277-786/13/$18 · doi: 10.1117/12.2008345

Good supervised learning algorithms are substantially more immune to spurious patterns in the data created by noise for the following reason: under the mostly reasonable assumption that the process contributing the noise operates independently from the class labels, the noise in the features will not correlate well with the class labels, and the algorithm will learn to ignore those patterns arising from noise. Unsupervised models, in contrast, have no grounding in labels to prevent them from confusing patterns that emerge by chance in the noise with the "true" patterns of potential interest. For example, even on clean data, LDA will often do poorly if the very simple feature selection step of removing stop-words is not performed first.

Though we call the models discussed here "supervised" topic models, it should be clarified that these models are only supervised in terms of the real-valued metadata. This is not supervision in the classical sense, where training data is supplied in the form of complete data, with values given for the latent variables in the target or evaluation data. The supervision in supervised topic models is much more akin to observing an additional feature for each document, a feature which may or may not be known for test data. The way in which words cluster by co-occurrence and in correlation with the metadata is still unsupervised. It is hoped that this extra feature will contribute information and improve the quality of the topics found by the model. Another advantage for most supervised topic models is their ability to be used to predict the metadata values for documents missing this information.

In this paper we show how the performance of supervised topic models degrades as character-level noise is introduced. Though we expect model quality to decrease in the presence of OCR errors, it is not well understood how sensitive supervised topic models are to such errors or how quality deteriorates as the level of OCR noise increases. We demonstrate the effect using both artificially corrupted data (which has several desirable properties due to our ability to choose the source data and the amount of corruption) and an existing real-world OCR corpus (with subsets from two separate decades) and measure model performance both in terms of the ability of the models to effectively predict missing metadata values and in terms of the quality of the topics discovered by the model. Because of the difficulties in evaluating topic models, even on clean data, these results should not be interpreted as definitive answers, but they do offer insight into prominent trends. It is our hope that this work will lead to an increase in the usefulness of collections of OCRed texts, as supervised topic models expose useful patterns to historians and other scholars.

The remainder of the paper is structured as follows: Section 2 discusses the most closely related work to this paper; Section 3 introduces the models that will be evaluated; Section 4 introduces the datasets used in our evaluation; Section 5 introduces the methodology of our evaluation including the metrics and experimental procedure employed; and Section 6 presents our conclusions and future work.

2. RELATED WORK

Both supervised and unsupervised topic models have been used previously to process documents digitized by OCR, including eighteenth-century American newspapers,[7] OCRed editions of *Science*,[8] OCRed research papers,[3] and books digitized by the Open Content Alliance.[9] Most of this previous work ignores the presence of OCR errors or attempts to remove corrupted tokens with special pre-processing such as stop-word removal and frequency cut-offs.

Similar evaluations to ours have been conducted to assess the effect of OCR errors on supervised document classification,[10,11] information retrieval,[12,13] and a more general set of natural language processing tasks.[14] Results suggest that in these supervised tasks OCR errors have a minimal impact on the performance of the methods employed, though it has remained unclear how well these results transfer to unsupervised methods.

More directly related, unsupervised document clustering and topic modeling have been evaluated in the presence of OCR errors.[6] It was found that both clustering and topic modeling suffer increasing performance degradation as word error rates (WER) increase. In the case of document clustering, simple feature selection can almost completely ameliorate the deleterious effects of the noise, not only improving performance evaluations at each noise level but also yielding performance at high word error rates that is fairly similar to the performance at low rates. In the case of topic modeling, feature selection improves performance but does not alter the shape of the quality degredation curve without feature selection.

3. MODELS

We chose to compare the performance of three different supervised topic models as part of our evaluation: Supervised LDA (sLDA),[2] Topics Over Time (TOT),[3] and Topics Over Nonparametric Time (TONPT).[5] The Dirichlet Multinomial Regression model[4] was not chosen because it is completely conditional on the metadata (e.g., no distribution is proposed for the metadata, but other variables depend on the values of the metadata) and is therefore not easily usable for metadata prediction. All of these models share a common LDA core, modeling documents as mixtures over topics and topics as mixtures over words. This LDA base has been extended in each case in order to jointly model document metadata with the topics and words. In the case of sLDA, metadata are modeled as per-document variables using a generalized linear model on the topic proportion vectors for each document together with a vector of linear coefficients and a variance parameter. TOT models metadata as per-word random variables (if there is only one metadata label for the document it is repeated for every word in the document) distributed according to per-topic Beta distributions. TONPT is based on TOT, but replaces the Beta distributions with Dirichlet Process Mixtures of Normals, which are Bayesian nonparametric density estimators. In addition, we use a baseline that is equivalent to unsupervised LDA during training; for prediction, a linear model is fit to the document topic proportions; we refer to this baseline as the PostHoc model.

4. DATA

For our evaluations we used a real OCR dataset and two synthetic OCR datasets derived from common datasets used in the document modeling and text analytics literature. In all cases, we used timestamps associated with the documents as the metadata. The real OCR dataset is based on the Legacy Tobacco Documents Library[†] which was compiled for the legal track of the 2006-2009 Text Retrieval Conferences from documents made public as part of various court cases against US tobacco companies. The documents were OCRed by the University of California at San Francisco. They span a wide range of time from the early 1900s to the early 2000s and are found in a similarly wide range of quality with respect to the fidelity of the OCR output. For our experiments we created two subsets of the data: one consisting of 5000 documents created from the year 1970 through the end of 1979 (Tobacco 70s) and another subset of 5000 documents from 1990 through the end of 1999 (Tobacco 90s). Though gold standard transcriptions of the documents are not available against which word error rates could be calculated, we use these two datasets to represent relatively high and low word error rates (respectively) with the assumption that documents produced in the 1990s will have been produced using higher quality printing and replication technologies and preserved for shorter periods of time, yielding higher quality document images and higher quality OCR output than those produced in the 1970s.

In addition to the real OCR data, we also used two synthetic OCR datasets that were created by rendering common text analytics datasets as document images, stochastically degrading the images to various levels and then OCRing the degraded images.[15] The datasets consisted of the LDC annotated portion of the Enron email corpus and the Reuters21578 corpus, both in uncorrupted form and at 5 levels of increasing average degradation. The synthetic datasets are useful for three reasons: first, because we know the source text, it is possible to calculate the average word error rate *exactly*; second, the same data are available at varying levels of degradation with word error rates (WER) ranging from very close to 0 to close to 50%, so the effects of increasing degradation on prediction and topic quality can be assessed independently from the underlying content; finally, the synthetic data have both timestamps and topical class labels which can be correlated with the topics found by the various models in order to assess topic quality. Both the Tobacco and synthetic datasets were lower-cased, had stopwords removed, and were also variously processed with a set of feature selection algorithms described in Section 5.

5. EXPERIMENTS

In order to assess the effects of OCR errors on the models studied here we conducted a series of experiments. In these experiments, supervised topic models were estimated with 100 topics on the real and synthetic datasets described above. A 20-round cross-validation scheme was used in which the model was trained on 90% of the data, sampled randomly without replacement each round, and 10% of the data were withheld for the evaluation. The ability of each model to predict the metadata values for the held-out documents given only their text content was calculated and recorded each round.

5.1 Metrics

Here we discuss the metrics that were used to evaluate the various models both in terms of their ability to predict missing metadata and to infer quality topics. We used the formulation of R^2 used by Blei and McAuliff[2] to assess topic quality:

$$R^2(\mathbf{t}, \hat{\mathbf{t}}) = 1 - \frac{\sum_d (t_d - \hat{t}_d)^2}{\sum_d (t_d - \bar{t})^2},$$

[†] `http://legacy.library.ucsf.edu`

where t_d is the actual metadatum for document d, $\hat{t}_d$ is the prediction and $\bar{t}$ is the mean of the observed t_ds. For linear models this metric measures the proportion of the variability in the data that is accounted for by the model. More generally, it is one minus the relative efficiency of the supervised topic model predictor to a predictor that always predicts the mean of the observed data points and can be negative. This metric is useful in cases where minimizing the sum squared error is desirable, but is problematic when the true distribution of the metadata is skewed or multimodal, as one can achieve relatively high R^2 scores in these cases by predicting values with very low likelihood. For example, choosing a point with near-zero density halfway between two modes of equal height can lead to a high R^2, even though the probability of the true value being close to that point is near zero.

Because of this deficiency of R^2, a second metric was used based on the generalized 0-1 loss.[16] It is the proportion of test instances that are within a distance of Δ from the true value:

$$\textit{Zero-One}(\mathbf{t}, \hat{\mathbf{t}}; \Delta) = \frac{1}{N} \sum_{d} \begin{cases} 1 \text{ if } |t_d - \hat{t}_d| < \Delta \\ 0 \text{ otherwise} \end{cases}$$

where N is the number of test instances, $\Delta = 0.01 \cdot (t_{max} - t_{min})$, and t_{max} and t_{min} are the maximal and minimal observed metadata values respectively. When it is important that predictions are very close to the true values at least some of the time the 0-1 loss is an appropriate metric.

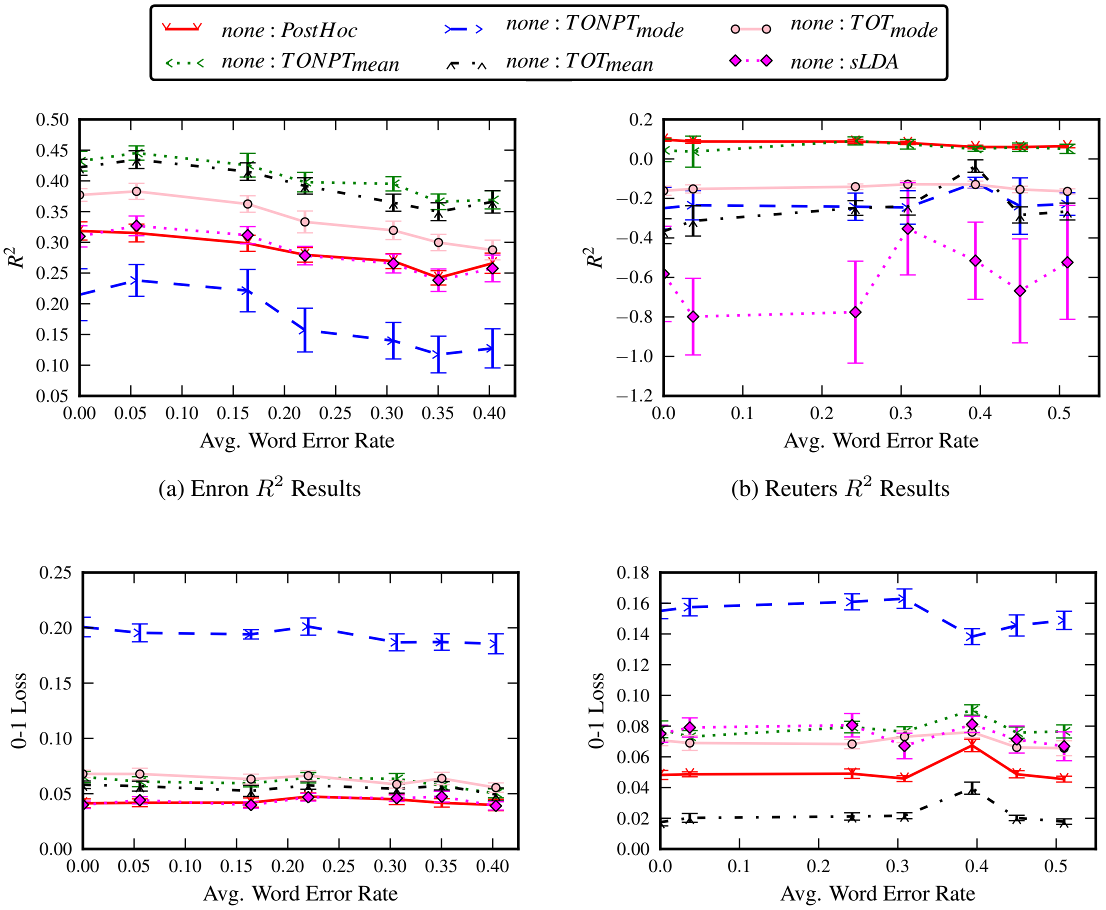
Topic quality was assessed using two metrics: half-document perplexity and an N-fold cross-validation metric (CV Accuracy). The half-document perplexity was calculated using a procedure similar to the one described by Rosen-Zvi et. al.[17] in which point estimates for the topic-word and document-topic categoricals distributions (for the test documents) were generated using the training data together with the metadata for the test documents and half of the words in each test document. Using these point estimates, the perplexity for the held-out words of the test documents was calculated.

The cross-validation metric is based on one first described by Griffiths et. al.[18] To compute this metric, the learned topic assignments are used as features in 10-fold cross-validation classification of the documents with the average accuracy across the folds defining the value of the metric. This evaluation mechanism avoids a potential problem that arises when evaluating topic models using a likelihood-based measure, such as perplexity, on noisy data where feature selection can significantly change the number of word types and tokens remaining in documents as the word error rate increases, giving the false impression that topic quality is greater than it really is.[6] The CV Accuracy metric is computed for each fold of the experiment (to be clear, that is 10 folds of cross-validated classification for each of the 20 folds of the larger topic modeling experiment). The means and standard errors of these four metrics across all twenty folds were recorded.

For prediction in the case of TOT and TONPT, we used two procedures discussed by Walker et al.[5] The first technique is based on the the one used by Wang and McCallum,[3] in which the posterior density for assigning a single value to all the per-token metadata variables of a test document (given the words in the test document and a proposed assignment of topic labels to those words) is calculated for a finite set of candidates and the candidate value with maximal posterior density is chosen. The topic labels were assigned using Gibbs sampling, treating the assignments made to the training data previously as given. The candidate metadata values were chosen by sampling values from the topic-conditional metadata distribution for each word given that word's topic assignment. We call this technique *mode prediction*, as it attempts to predict the value at the (maximal) mode of the joint posterior of the metadata given the words in the document. The second prediction procedure also made use of sampled topic assignments. The assignments were used to estimate the document-specific distribution over topics, and this distribution was used to calculate the expected value of the document metadata variables as a weighted average of the expected value of the topic-specific metadata distributions. We call this technique *mean prediction*, as it attempts to predict the mean or expected value of the posterior metadata distribution.

5.2 Feature Selection

To determine the degree to which the effects of OCR errors could be mitigated, we experimented both on "raw" OCR documents and on OCR documents processed using various unsupervised feature selection algorithms. The first feature selector was a term frequency cut-off filter (TFCF), with a cut-off of 5 as used by Wang and McCallum[3] (indicated with `tfcf.5` in the plots). The next selector was a proportion filter (`proportion`) which removes any word occurring in fewer than 1% of the documents or in more than 50%. The next selector was Term Contribution (TC), originally developed for document clustering[19] and parameterized by the number of word types that are to remain after selection. We attempted two values for this parameter: 10K and 50K (`tc.10000`, and `tc.50000`). The final method we employed was Top-N per



(a) Enron R^2 Results

(b) Reuters R^2 Results

(c) Enron 0-1 Loss Results

(d) Reuters 0-1 Loss Results

Figure 1: Timestamp prediction results for the two synthetic datasets without feature selection

Document (TNPD),[20] which first assigns each word type in every document a document-specific score (e.g., its TF-IDF weight) and then selects words to include in the final vocabulary by choosing the N words with the highest score from each document (TNPD with $N = 1$ is abbreviated as `tnpd.1`). In many of the plots and graphs that follow, the feature selector that was used is specified before the name of the algorithm, separated by a colon.

5.3 Synthetic Dataset Results

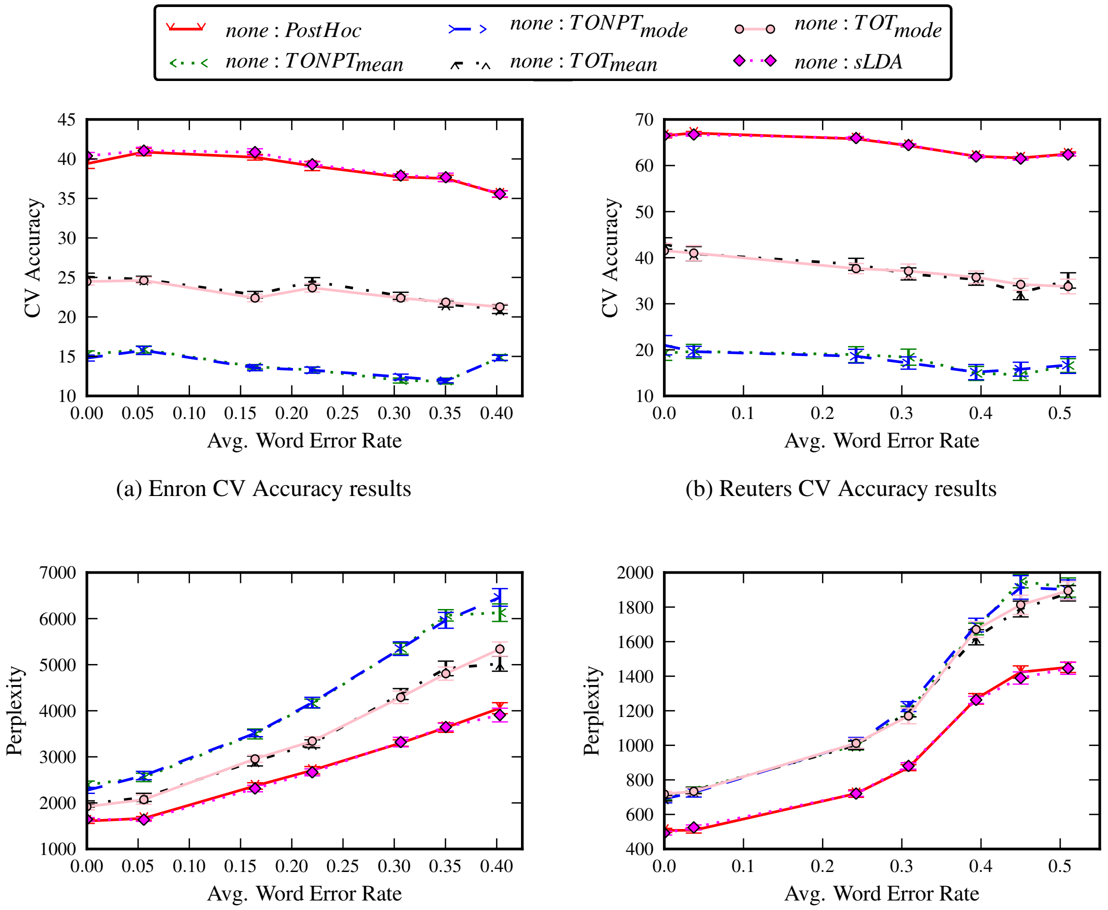
Here we discuss the results of the experiments on the raw (without feature selection) synthetic data.

5.3.1 Metadata Prediction Quality

Figure 1 shows the prediction results for the raw data. In terms of metadata prediction, Figure 1 shows that the algorithms appear to produce a fairly wide range of outcomes according to the R^2 metric. For the Enron data the TONPT and TOT models have a slight edge when used together with the mean prediction algorithm, although all of the curves appear to trend downward. In the case of the Reuters data, TONPT is mostly tied with the PostHoc baseline, with the other models performing mostly worse than the TONPT mean predictor. As the text error rate increases, there also appears to be a downward trend for the Reuters data, though the variance in the results is greater and the trend is difficult to discern. With respect to the 0-1 Loss metric, the TONPT model with the mode prediction algorithm is clearly superior to the other models.

5.3.2 Topic Quality

According to the two topic quality metrics, shown in Figure 2 the methods fare quite differently. Both the PostHoc baseline and sLDA are clearly superior in this regard, though their respective performance is indistinguishable. The TOT model fares



(a) Enron CV Accuracy results

(b) Reuters CV Accuracy results

(c) Enron Half-document Perplexity results

(d) Reuters Half-document Perplexity results

Figure 2: Topic quality results for the two synthetic datasets without feature selection (lower perplexity is better).

significantly worse and the TONPT model worse still. In addition, the graphs seem to suggest that all of the models degrade in performance at roughly the same rate, as the shapes of the curves in the graphs are fairly similar. This is especially obvious in the Reuters results, where the distinct difference in performance from the 0.3 to 0.4 word error rates is mirrored in similar upticks (for perplexity) and dips (for the cross-validation metric) across all of the models.

5.3.3 Significance of Trends

In order to analyze the effects of noise more objectively, we attempted to quantify the effects of noise on performance by testing whether the results at each WER level were statistically significantly worse than results on the clean un-degraded data. This was determined using a one-sided stochastic permutation test of the hypothesis that the mean of the results at a given word error rate was worse (greater than in the case of perplexity, less than in the case of the other metrics) than the mean results on the clean data. P-values less than .05 were considered significant. Because of the large number of tests conducted here, we control for likely Type I errors by only considering a result significant if the results at all higher WERs are also significant or, in the case of the highest WER, if the the P-value is less than 0.001. Figure 3 shows, for each dataset and each metric, at what lowest WER a significantly worse outcome was found. These results suggest that all of the algorithms experience significant degradation in performance as the WER increases, often even at relatively low WERs. In some cases, the supervised topic models appear to be slightly more robust, degrading at higher WERs than the PostHoc baseline, although that is not always true: several of the supervised topic models experience degradation at lower error rates in term of half-document perplexity and CV Accuracy on the Reuters dataset than the baseline. In other words, supervised topic models may have an advantage in terms of how topic quality degrades as error rates increase, but it is not a large one and appears inconsistently across datasets. Note that, though the dashes indicate that no significant degradation in

		Lowest WER with Statistically Significantly Worse:			
Dataset	**Model**	$\mathbf{R^2}$	**0-1**	**Perplexity**	**CV Accuracy**
Enron	PostHoc	0.1641	-	0.0556	0.3063
	sLDA	0.2200	-	0.1641	0.2200
	TOT_{mean}	0.2200	-	0.0556	0.1641
	TOT_{mode}	0.2200	0.4031	0.0556	0.1641
	$TONPT_{mean}$	0.2200	0.3503	0.0556	0.3063
	$TONPT_{mode}$	0.2200	0.3063	0.0556	-
Reuters	PostHoc	0.3084	-	0.2422	0.2422
	sLDA	-	-	0.0372	0.2422
	TOT_{mean}	-	-	0.0372	0.0372
	TOT_{mode}	-	-	0.2422	0.2422
	$TONPT_{mean}$	-	-	0.0372	0.3938
	$TONPT_{mode}$	-	-	0.0372	0.2422

Figure 3: The lowest word error rate at which each model had significantly worse performance than the same model on the "clean" data and all following WERs were also significantly worse. A dash indicates that none of results at higher WERs were worse than the results on the clean data, or if they were, there was an insignificant diference at a higher WER.

performance was found, they usually occur in combinations that have very poor performance to begin with. For example, the PostHoc and sLDA methods both did not experience significant degradation for the Enron dataset and the 0-1 Loss metric, but that is most likely because they are approaching random performance (See Figure 1c).

5.3.4 Feature Selection

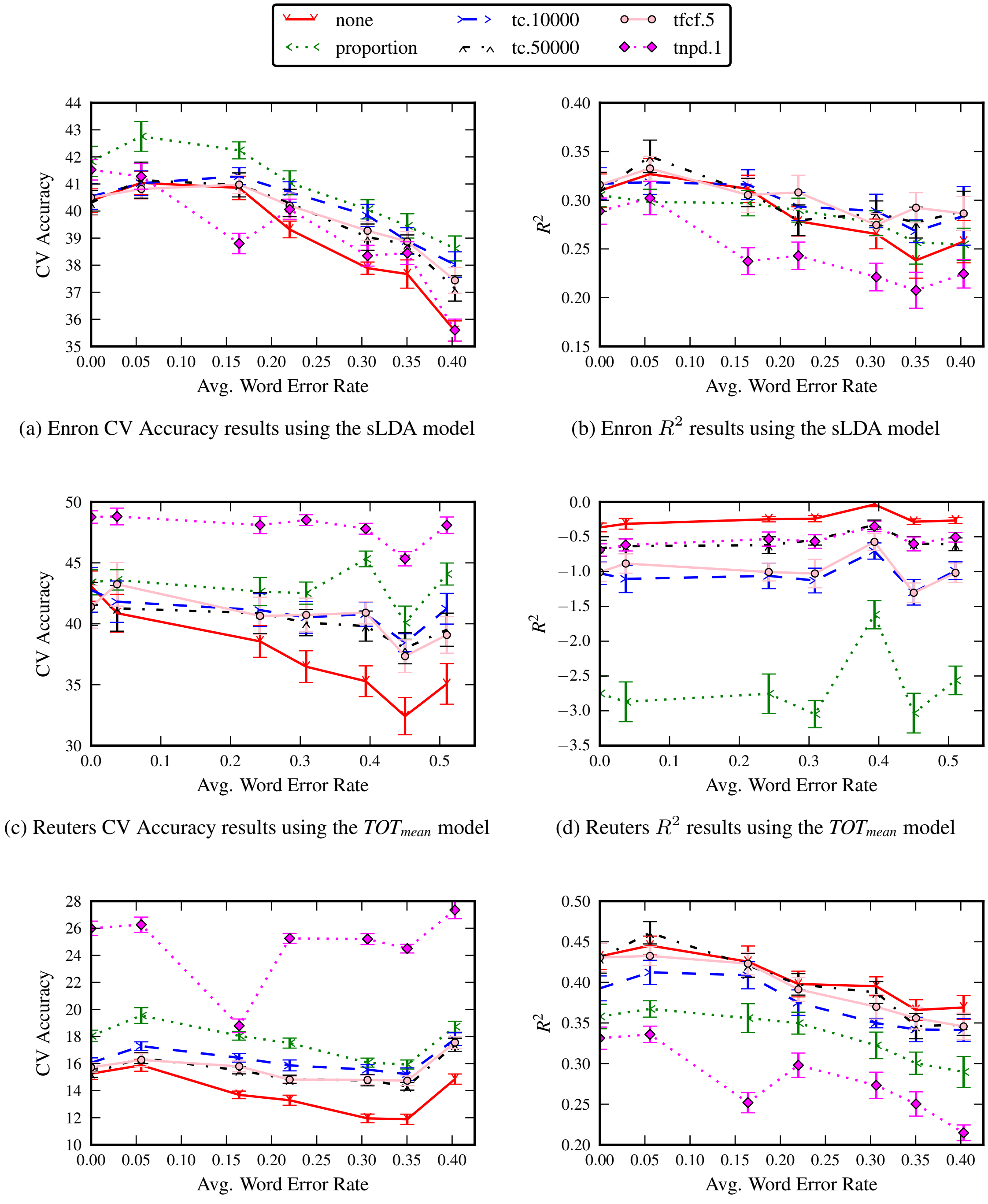
We wished to examine whether and to what extent unsupervised feature selection algorithms can ameliorate the deleterious effects of OCR noise on supervised topic models. We repeated the above experiments on each of the feature-selected versions of the synthetic data discussed above. Figure 4 shows a few representative examples of the types of trends we observed across all models and feature selection algorithms. Figure 5 shows how the models compare given a single feature selection algorithm (tnpd.1). We found that it was typical for feature selection to improve CV Accuracy values, but at the same time hurt the R^2 and 0-1 Loss values. Furthermore, the feature selectors that improve CV Accuracy the most (typically the TNPD and proportion methods) were the ones that hurt R^2 and 0-1 Loss performance the most.

These outcomes have an intuitive explanation if we consider the supervised topic modeling task as being composed of two somewhat orthogonal tasks: learning topic word clusterings and learning word metadata clustering. The learning of the topic clusters is an unsupervised task while the learning of the metadata clusters is supervised (since the metadata are observed for the training data). It has long been known that for supervised learning tasks, such as text classification, feature selection often hurts the learner's performance.[21] As discussed in Section 1, this is because a supervised algorithm is able to learn the correspondence between the features and the labels and gain information even from features that appear less correlated with the classifications a-priori. In contrast, feature selection is often essential in unsupervised learning tasks (e.g., document clustering and topic modeling[6]). This is because unsupervised learning algorithms have no frame of reference to distinguish extraneous patterns in the data from those that matter to the human conducting the analysis. So, in the case of supervised topic models, feature selection has the natural consequence of helping the performance of the unsupervised learning facet of the task and harming the performance of the supervised facet.

5.4 Real Dataset Results

As in the case of the synthetic datasets, we show results without and with feature selection. The results for the Tobacco 70s and Tobacco 90s datasets without feature selection are shown in Figure 6. CV Accuracy results were not calculated for the Tobacco data because it lacks topical labels for the documents.

The results match the findings on the synthetic dataset without feature selection in terms of the trends across models and across noise levels. In the case of each of the models across the evaluations used, the performance was better for the data from the 1990s. It is possible that some of the difference in performance could be attributed to variables other than increased noise; for example, it could be the case that the environment of the 1970s was more static and that less changed from year to year, making timestamp prediction based solely on words difficult. This seems unlikely, though, since not only the predictions, but also the topic quality (as measured with half-document perplexity) are impacted.



(a) Enron CV Accuracy results using the sLDA model

(b) Enron R^2 results using the sLDA model

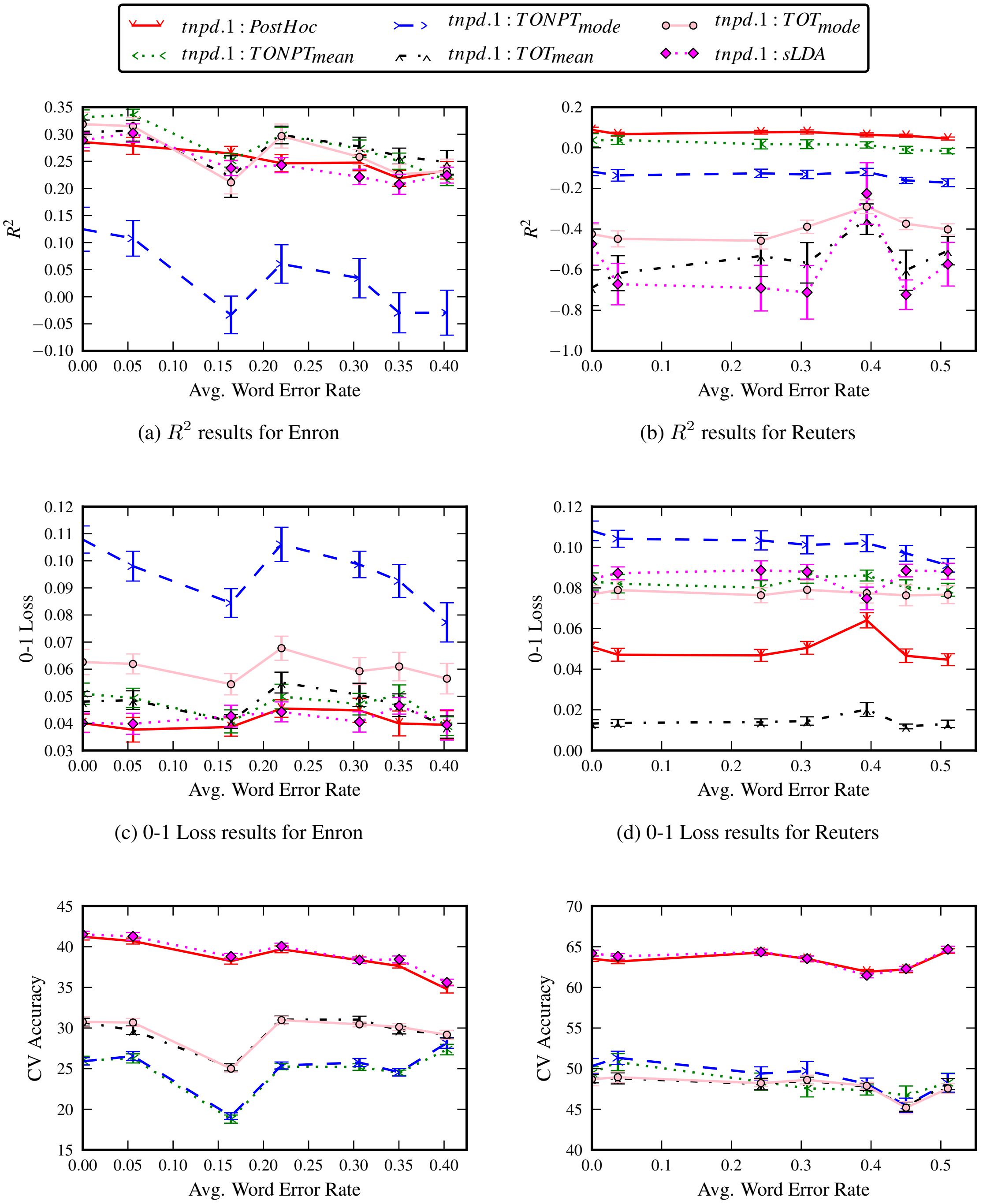
(c) Reuters CV Accuracy results using the TOT_{mean} model

(d) Reuters R^2 results using the TOT_{mean} model

(e) Enron CV Accuracy results using the $TONPT_{mean}$ model

(f) Enron R^2 results using the $TONPT_{mean}$ model

Figure 4: Example results for three (dataset, selector) pairs showing the effects of the feature selectors on CV Accuracy and R^2 scores. (a) and (b) show results on Enron with sLDA. The proportion and TFCF selectors improve both metrics, especially at higher error rates. (c) and (d) show the results on Reuters using TOT_{mean}. TNPD and proportion improve CV Accuracy, but all of the selectors hurt R^2 scores. (f) and (e) show the results on Enron using $TONPT_{mean}$. TNPD and proportion improve CV Accuracy, but all of the selectors hurt R^2 scores.



(a) R^2 results for Enron

(b) R^2 results for Reuters

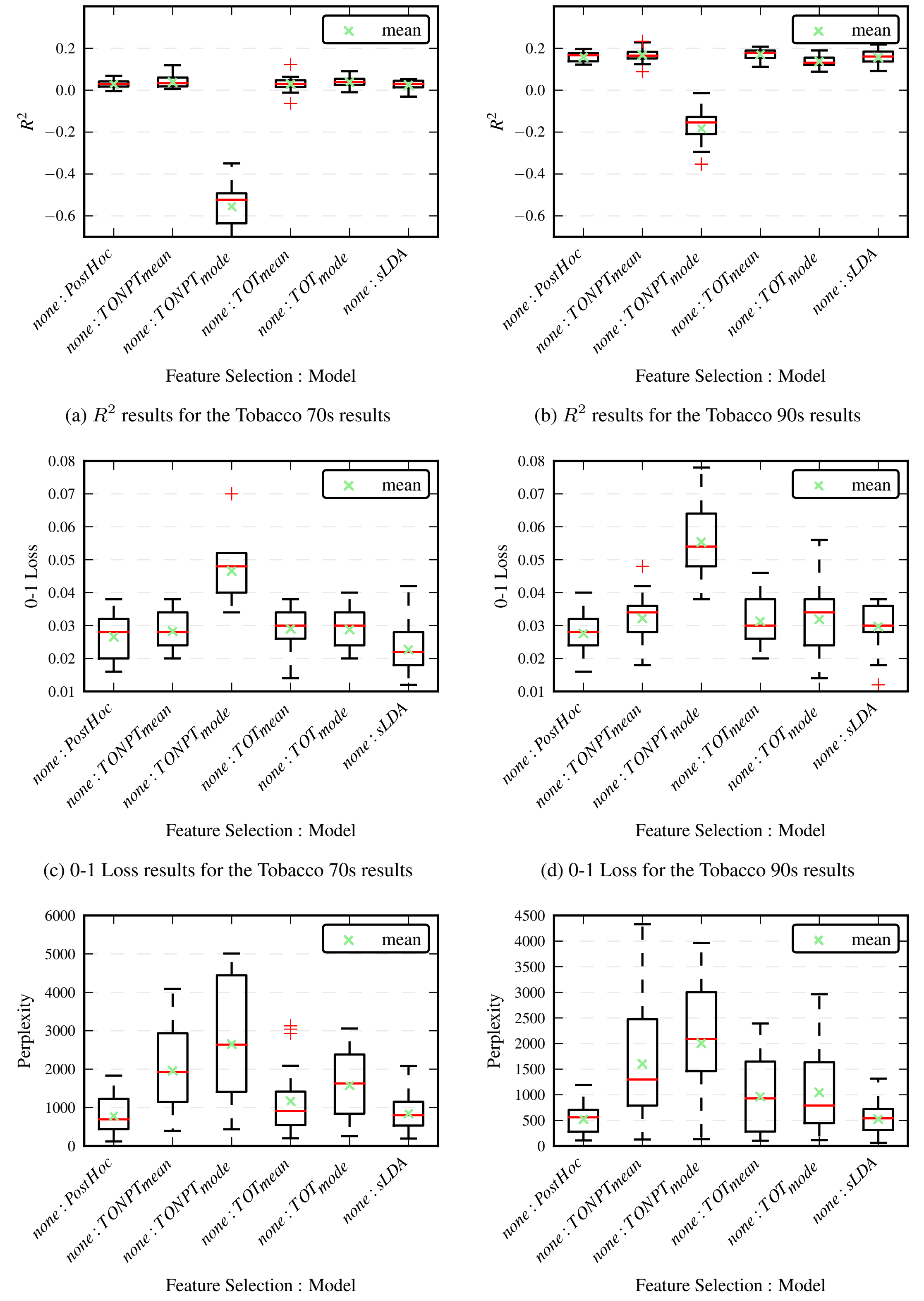
(c) 0-1 Loss results for Enron

(d) 0-1 Loss results for Reuters

(e) Cross Validation Accuracy results for Enron

(f) Cross Validation Accuracy results for Reuters

Figure 5: Results for the two synthetic datasets with TNPD feature selection. A comparison of these plots to those in Figures 1 and 2 show trends in the effect that feature selection has on the performance of the models. Specifically, while topic quality is improved (as evidenced by generally higher CV Accuracies), metadata prediction performance is actually hurt (according to both the R^2 and 0-1 Loss metrics). Half-document perplexity is not shown because feature selection skews that metric making it unreliable.



(a) R^2 results for the Tobacco 70s results

(b) R^2 results for the Tobacco 90s results

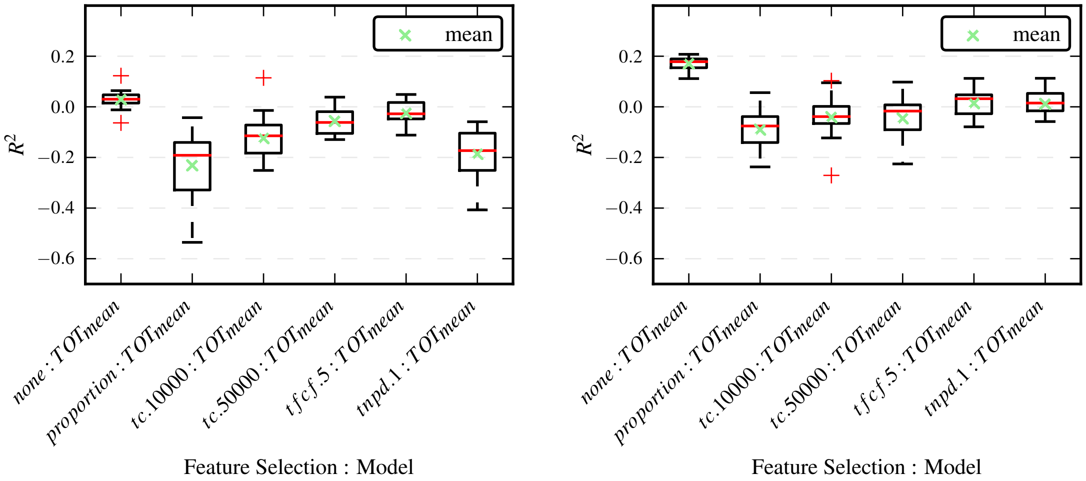
(c) 0-1 Loss results for the Tobacco 70s results

(d) 0-1 Loss for the Tobacco 90s results

(e) Perplexity for the Tobacco 70s results (lower is better)

(f) Perplexity for the Tobacco 90s results (lower is better)

Figure 6: Results for the two real world datasets without feature selection.



(a) R^2 results for the Tobacco 70s results

(b) R^2 results for the Tobacco 90s results

Figure 7: R^2 results for the two real world datasets with feature selection using the TOT_{mean} model. All of the feature selectors result in worse performance than is achieved without feature selection (none).

Ftr. Selection	none	tnpd.1	none	tnpd.1	none	tnpd.1	none	tnpd.1
Words	alveolar	alveolar	cholesterol	blood	paper	filter	sales	coupon
	cells	lung	patients	cholesterol	mm	mm	market	promotion
	pulmonary	cells	subjects	patients	filter	paper	brand	sales
	bacteria	macrophages	blood	heart	tipping	plug	promotion	carton
	lung	tissue	ldl	coronary	plug	tipping	total	pack
	macrophages	bacteria	serum	regan	weight	weight	year	display
	bacterial	pgnbr	disease	serum	wrap	dilution	advertising	coupons
	lymphatic	blood	cv	disease	length	acetate	share	purchase
	tissue	particles	age	uptake	pressure	wrap	coupon	retail
	blood	cell	healthy	plasma	cigarette	drop	media	brand

Figure 8: The top 10 most probable words from four pairs of similar topics produced by the sLDA algorithm on the Tobacco 70s dataset with no feature selection and with TNPD feature selection.

Again, we repeated the experiments after processing the Tobacco datasets with the feature selection algorithms. Figure 7 shows a result that was typical across the various supervised topic models. On this data, like on the synthetic data, feature selection appeared to mostly harm prediction results.

It is more difficult to assess the impact of feature selection on topic quality for the Tobacco datasets, as there are no human-supplied topic labels. A visual inspection of the most probable words in topics produced with and without feature selection did not reveal substantial differences in topic quality. Figure 8 shows similar topic pairs found on two runs of the sLDA algorithm on the Tobacco 70s dataset, one without and one with TNPD feature selection. The top 10 most probable words for each topic are listed. Although there are differences between the similar topic pairs, it is difficult to claim that any are significantly better than their counterparts.

6. CONCLUSION AND FUTURE WORK

We assessed the impact of OCR errors on the performance of the supervised topic models Supervised LDA, Topics Over Time, and Topics Over Nonparametric Time. Our results indicate that, despite having more information about the data, supervised topic models do not seem to be more robust to noise than traditional topic models. As document quality decreased, topic quality decreased at roughly the same pace for the supervised topic models as for the traditional topic model (PostHoc) baseline. Prediction quality also degraded significantly with increased error rates in the majority of cases. Feature selection improved the performance of the models in terms of topic quality, but at the same time it hurt the performance of the models in terms of metadata prediction. It is possible that a feature selector tailored to this task might help alleviate this problem.

For example, one might use information gain or distributional mutual information to select word features that correlate well with both the topics and metadata, thereby increasing topic quality while not greatly decreasing prediction quality.

Acknowledgments

We acknowledge the Fulton Supercomputing Center at BYU for providing the computing resources required for this research.

REFERENCES

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent Dirichlet allocation," *Journal of Machine Learning Research* **3** (2003).

[2] Blei, D. M. and McAuliffe, J. D., "Supervised topic models," *arXiv:1003.0783* (Mar. 2010).

[3] Wang, X. and McCallum, A., "Topics over time: A non-Markov continuous-time model of topical trends," in [*Proceedings of the 12th ACM SIGKDD International Conference*], (Aug. 2006).

[4] Mimno, D. and McCallum, A., "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression," in [*Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*], AUAI Press (2008).

[5] Walker, D. D., Seppi, K., and Ringger, E. K., "Topics over nonparametric time: A supervised topic model using Bayesian nonparametric density estimation," in [*Proceedings of the 9th Bayesian Modelling Applications Workshop*], (Aug. 2012).

[6] Walker, D. D., Lund, W. B., and Ringger, E. K., "Evaluating models of latent document semantics in the presence of OCR errors," in [*Proceedings of the Conference on Empirical Methods in Natural Language Processing*], (2010).

[7] Newmann, D. J. and Block, S., "Probabilistic topic decomposition of an eighteenth-century American newspaper," *Journal of the American Society for Information Sciences and Technology* **57** (Feb. 2006).

[8] Blei, D. M. and Lafferty, J. D., "Dynamic topic models," in [*Proceedings of the 23rd International Conference on Machine Learning*], (June 2006).

[9] Mimno, D. and McCallum, A., "Organizing the OCA: learning faceted subjects from a library of digital books," in [*Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*], ACM Press (2007).

[10] Taghva, K., Nartker, T., Borsack, J., Lumos, S., Condit, A., and Young, R., "Evaluating text categorization in the presence of OCR errors," in [*Proceedings of the IS&T/SPIE 2001 International Symposium on Electronic Imaging Science and Technology*], SPIE (2001).

[11] Agarwal, S., Godbole, S., Punjani, D., and Roy, S., "How much noise is too much: A study in automatic text classification," in [*Proceedings of the 7th IEEE International Conference on Data Mining*], (2007).

[12] Taghva, K., Borsack, J., and Condit, A., "Results of applying probabilistic IR to OCR text," in [*Proceedings of the 17th International ACM/SIGIR Conference on Research and Development in Information Retrieval*], (1994).

[13] Beitzel, S. M., Jensen, E. C., and Grossman, D. A., "A survey of retrieval strategies for OCR text collections," in [*In Proceedings of the Symposium on Document Image Understanding Technologies*], (2003).

[14] Lopresti, D., "Optical character recognition errors and their effects on natural language processing," in [*Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*], ACM (2008).

[15] Walker, D. D., Lund, W. B., and Ringger, E. K., "A synthetic document image dataset for developing and evaluating historical document processing methods," in [*Proceedings of Document Recognition and Retrieval XIX*], (Jan. 2012).

[16] Pratt, J. W., Raïffa, H., and Schlaifer, R., [*Introduction to Statistical Decision Theory*], MIT Press (1995).

[17] Rosen-Zvi, M., Griffiths, T. L., Steyvers, M., and Smyth, P., "The author-topic model for authors and documents," in [*Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*], AUAI Press (2004).

[18] Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B., "Integrating topics and syntax," in [*Advances in Neural Information Processing Systems 17*], Saul, L. K., Weiss, Y., and Bottou, L., eds., **17**, MIT Press (2005).

[19] Liu, T., Liu, S., Chen, Z., and Ma, W., "An evaluation on feature selection for text clustering," in [*Proceedings of the 20th International Conference on Machine Learning*], (Aug. 2003).

[20] Walker, D. D. and Ringger, E. K., "Top N per document: Fast and effective unsupervised feature selection for document clustering," Tech. Rep. 6, Brigham Young University (July 2010). `http://nlp.cs.byu.edu/techreports/BYUNLP-TR6.pdf`.

[21] McCallum, A. and Nigam, K., "A comparison of event models for naive Bayes text classification," in [*Proceedings of the AAAI-98 Workshop for Text Categorization*], (July 1998).