We thank the reviewers for their time and careful reviews. Please find our responses below.

General. We will update the paper with NPMI topic coherence scores and compare to results published by Miao et al. (2018). We find that logistic LDA outperforms other *topic models*, but performs similar to *document models* which represent topics implicitly and are thus more difficult to interpret.

	$K = 50$	$K = 200$	Explicit topics	Generative
LogisticLDA	0.215	0.179	✓	×
OnlineLDA	0.131	0.112	✓	✓
NVLDA	0.110	0.110	✓	✓
NVDM	0.186	0.157	×	✓
ProdLDA	0.240	0.190	×	✓

Reviewer 1. The relationship between sLDA and logistic LDA is discussed in Section 3.1 and illustrated in Figure 1. We will add some detail on discLDA and LLDA. As for empirical comparisons with sLDA, we tried to get its open-source code to work on our supervised tasks, but it did not scale to the size of the Twitter dataset. We thus extended another implementation of LDA to be supervised in the same manner as logistic LDA (Eq.6). What we call "LDA" in Table 1 is in fact this supervised extension. This is related to your comment regarding the necessity to better explain how we extended the implementation of Theis and Hoffman (2015). We appreciate this remark and will improve our paper on this front.

Like LDA, logistic LDA indeed discovers a distribution over topic proportions for each document (Eq. 1 and 12), as well as over topics for each item (Eq. 2 and 13). We will make this point more explicit by improving lines 175-178.

Regarding the Twitter dataset, we will release lists of tweet IDs annotated by topic IDs. We will not provide the correspondence between topic IDs and topic names or the tweet text itself to comply with legal policies such as GDPR.

Where authors are annotated with multiple communities/topics in the dataset, we chose one at random. This means the ground-truth label is noisy, limiting the maximum achievable accuracy.

Reviewer 2. We are grateful for the provided references and will include them in Section 2.2. We will also add more detail on other supervised extensions.

Regarding the "alternative view" of LDA, we appreciate the feedback and will try to improve its presentation. Note that the assumptions in Eq. 1 to 3 together with Eq. 4 fully characterize LDA, where instead of deriving LDA from priors and likelihoods, we derive LDA from a set of conditional distributions. This view suggests other modifications to the classical view of LDA, and our paper presents one of them. Future work may want to explore alternatives to the Dirichlet assumption, but this is beyond the scope of our paper.

Regarding the loss of beauty and probabilistic foundations, note that probabilistic models do not have to be generative but they are merely sets of assumptions. LDA corresponds to one such set of assumptions and logistic LDA to a modified set of assumptions, which in both cases yield a proper probabilistic model.

In the Twitter and 20-Newsgroups experiments, having a one-to-one mapping from topics to classes seems to be a reasonable assumption, but we agree that in general, more flexibility is desirable. To relax this assumption, we outlined a solution in lines 137-139. In defense of our title, please note that it does not mention that our model is supervised. Logistic LDA is a discriminative topic model in the sense that it does not model the distribution of inputs (see also lines 301-305). Logistic LDA can be unsupervised, supervised or semi-supervised. Regarding the appendix, it may appear lengthy because we included derivations of some known results so that readers do not have to consult multiple sources.

It is indeed incoherent to use *accuracy* in Table 1 and *error rate* in Table 2. We will report accuracy in both tables.

The reported training time of 11 hours was obtained using a CPU. It reduces to less than 1 hour when using a GPU, which is considerably faster compared to LSTM-based models. An SVM is even faster, however, its results are not competitive. We will update the appendix with GPU training time estimates.

Reviewer 3. Regarding extensibility, logistic LDA can be modified in the same ways as LDA when the changes target common components of the two models, such as the distribution of topics π_d. For example, one could replace the Dirichlet with a logistic normal distribution or allow for an unbounded number of topics. Our supervised version of logistic LDA illustrates that adding extra information to the model and using different losses is as easy as in LDA.

We implemented both online and batched versions of training and inference algorithms. In the former case, the updates can be done after seeing a single word or a batch of words from possibly multiple documents. This allows us to process large collections of arbitrary length, which is particularly interesting for the Pinterest and Twitter cases, where more pins or tweets become available. We will add this discussion on scalability to Section 5.1.