Re-Implementation of *Latent Dirichlet Allocation*

August Regnell	**Joar Haraldsson**	**Emil Kerakos**	**Javier García San Vicente**
970712-9491	951111-1230	960720-7777	981027-T131
aregnell@kth.se	joarha@kth.se	ekerakos@kth.se	jgsv@kth.se

15 January, 2021

Abstract

The aim of this paper was to re-implement the generative probabilistic model *latent Dirichlet allocation* (LDA) proposed by David M. Blei, Andrew Y. Ng, and Michael I. Jordan, to replicate a subset of their experiments and to extend on their work.

Approximate inference techniques and an EM-algorithm was implemented for parameter estimation in the smoothed and unsmoothed LDA-algorithm, and the same data sets as in the original LDA paper have been used to ensure results being comparable.

In terms of text classification, our results, albeit not identical to those found by Blei et al., support that LDA maintains much discriminatory information during dimensionality reduction. The discrepancy in achieved results, are thought to stem from different data preprocessing. Meanwhile, our analysis of collaborative filtering show that unsmoothed LDA, with a simple heuristic, outperformed a mixture of unigram model for low numbers of topics, while smoothed LDA did not yield similar results as found by Blei et al.

Furthermore we used the algorithm to estimate how the amount of media attention given to certain topics changed over time in a digital newspaper corpus. Without employing dynamic topic models the results obtained we're deemed to have support in the timing of actual news events and reporting.

We conclude that LDA effectively models latent topics of different kinds of discrete data, such as text corpora and movie user data. It can also easily be used for classification, trend modelling and recommendation. However, the model is computationally heavy and inflexible regarding refitting to new data. Later research has extended the model to combat these problems, but the original idea still remains one of the most popular within topic modelling.

1 Introduction

1.1 Latent Dirichlet Allocation and Topic Modelling

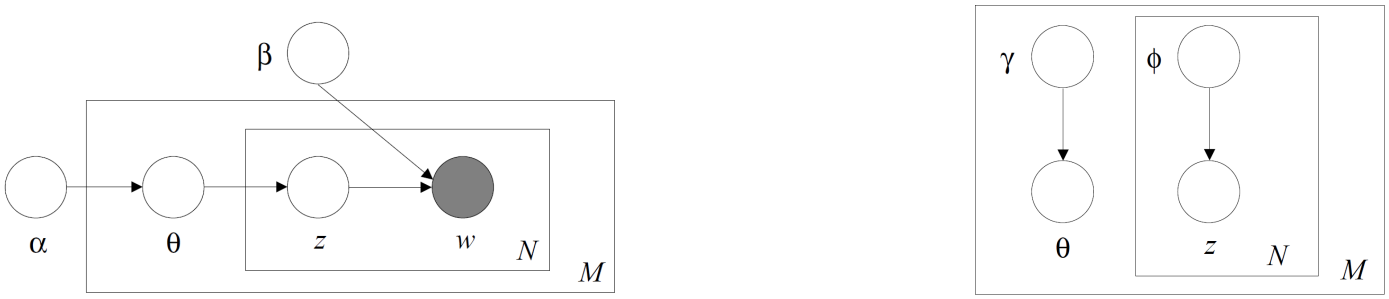
In 2001 David M. Blei, Andrew Y. Ng, and Michael I. Jordan proposed Latent Dirichlet Allocation (LDA): a probabilistic generative model for collections of discrete data [3]. The model was developed in large for purposes of topic modelling in large corpora where the aim is to find low-dimensional representations of documents – or more generally, members – in a collection while retaining statistical relationships useful for tasks relating to information retrieval [3]. Essentially the goal is to automatically discover the thematic structure in a large corpus and "annotate" each document with the topics present in the document [1]. For instance if an article (the document) in a newspaper archive (the corpus) is about an injury a soccer player has sustained in a game and the consequential economic loss of the player getting injured in a contract year: the topic modelling algorithm might then annotate the document with topics "sports", "health" and "wealth". These topic representations of documents can then be used to summarize, organize and make large collections of text searchable, but they can also be used for other IR-tasks such as novelty detection and classification [3] [1].

1.2 LDA - the model

Being a probabilistic generative model, LDA assumes that each document m is generated in the following way (depicted in figure 1a).

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n

α is the parameter of the Dirichlet prior of the distribution of topics for every document. β parameterizes the word probabilities where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$. θ_d is the distribution of the topics in document d, informally the "topic proportions" of the document. z_{dn} is the topic of the n:th word in document d, w_{dn}.



(a) (b)

Figure 1: Graphical models for (a) the full LDA model and (b) the variational distribution used to approximate the posterior in LDA. The outer plate represents the documents, the inner plates the topics and words within a document.

1.3 Inference and parameter estimation

With the model in place, the authors proposed approximate methods for performing probabilistic inference and parameter estimation, given a corpus of documents.

Variational Inference of document-level and word-level variables

To find short descriptions and discover the thematic structure of a document one wishes to infer the posterior distribution of the hidden document-level variable θ (the topic proportions in the document) and the hidden word-level variables $\mathbf{z}$ (the hidden topic of each word in the document). With the assumptions above the true posterior is given by:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)} \tag{1}$$

However, this distribution is intractable to compute in general due to the coupling between θ and β. Indeed,

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \tag{2}$$

One can instead consider a simpler model as presented in figure 1b. This is a simplification of the original LDA model where some edges and nodes have been removed. Since the edges between θ, $\mathbf{z}$ and $\mathbf{w}$ have been dropped we get free variational parameters and obtain a family of distributions on the latent variables which are characterized as follows:

$$q(\theta_d, \mathbf{z}_d \mid \gamma_d, \phi_d) = q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(z_{dn} \mid \phi_{dn}) \tag{3}$$

where γ_d is the Dirichlet parameter and $(\phi_{d1}, \ldots, \phi_{dN})$ are the multinomial parameters. These parameters are now free variational parameters and we can now determine them using Variational Inference (VI).

We want to minimize the Kullback-Leibler divergence between the variational distribution and the true posterior, $D(q(\theta, \mathbf{z} \mid \gamma, \phi) \mid\mid p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta))$, to get a good approximation. This is equal to maximizing the evidence lower bound, L_d, for every document, which is equal to:

$$L_d(\gamma, \phi; \alpha, \beta) = \mathbb{E}_q[\log p(\theta \mid \alpha)] + \mathbb{E}_q[\log p(\mathbf{z} \mid \theta)] + \mathbb{E}_q[\log p(\mathbf{w} \mid \mathbf{z}, \beta)] - \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log q(\mathbf{z})] \tag{4}$$

Expanding the expectations and putting the partial derivative of L_d w.r.t. the variational parameters we get the following update equations.

$$\phi_{dni} \propto \beta_{iw_n} \exp\left\{\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k} \gamma_{dj})\right\}, \qquad \gamma_{di} = \alpha_i + \textstyle\sum_{n=1}^{N} \phi_{dni}$$

where Ψ is the digamma function, that is the first derivative of the $\log \Gamma$ function.

Parameter estimation of corpus-level parameters

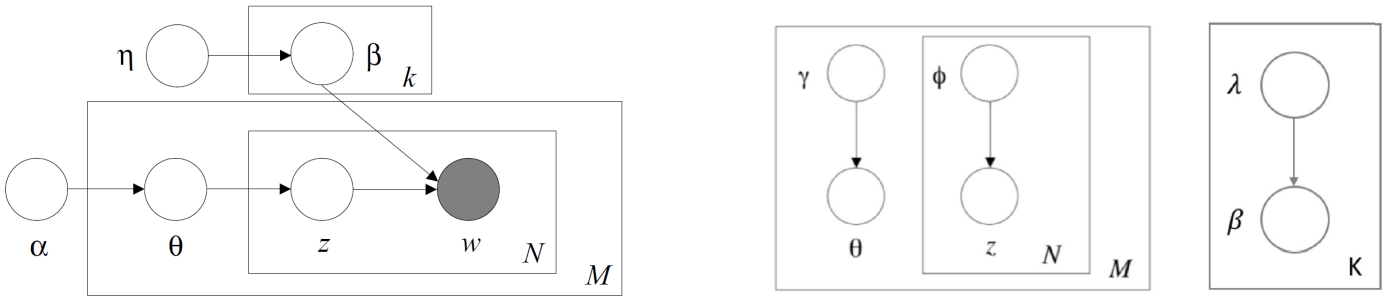
To estimate the corpus-level parameters, α and β, we add all document-level lower bounds and maximize w.r.t. to α and β.

The update for the Dirichlet parameter α is calculated using an efficient Newton-Raphson (NR) method (note that there are a couple errors in the derivation in the original paper). The update for the conditional multinomial parameter β can be calculated analytically. We get the following expression:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn} \tag{5}$$

We can now construct an EM-algorithm. In the E-step γ^*_d and ϕ^*_d are optimized for every document d with the VI-algorithm. In the M-step α and β are updated. The algorithm is finished when the lower bound has converged.

Bayesian modelling of corpus-level parameters (smoothing)



(a) (b)

Figure 2: Graphical models for (a) the full smoothed LDA model and (b) the variational distribution used to approximate the posterior in smoothed LDA.

Due to document corpora usually having large vocabulary sizes one often runs in to serious problems regarding sparsity. This can be combatted with smoothing, but the usual methods are not justified in the mixture model setting, and we thus use a different method. One can instead apply variational methods to the extended model in figure 3 that includes an exchangeable Dirichlet (parameter η) on each row in β. We now get

$$q(\beta_{1:k}, \mathbf{z}_{1:M}, \theta_{1:M} \mid \lambda, \phi, \gamma) = \prod_{i=1}^{k} \mathrm{Dir}(\beta_i \mid \lambda_i) \prod_{d=1}^{M} q_d(\theta_d, \mathbf{z}_d \mid \phi_d, \gamma_d) \tag{6}$$

where q_d is the variational distribution defined in the un-smoothed model. The corpus-level lower bound, $L = \sum_{d=1}^{M} L_d$, now gets two additional terms $\mathbb{E}_q[\log p(\beta \mid \eta)]$ and $-\mathbb{E}_q[\log q(\beta)]$. Maximizing in the same way as before, we

get the same equation for γ_d, but new equations for ϕ and λ, namely:

$$\phi_{dni} \propto \exp\Big\{\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj}) + \Psi(\lambda_{iw_n}) - \Psi(\sum_{l=1}^{V}\gamma_{il})\Big\}, \qquad \lambda_{ij} = \eta + \textstyle\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi^*_{dni}w^j_{dn}$$

α can be found using the same Newton-Raphson algorithm as before. The hyperparameter η can be found using a similar version of the same algorithm, but since it's a scalar it becomes simpler. We get $\eta_{new} = \eta_{old} - \frac{\partial L}{\partial \eta}\big(\frac{\partial^2 L}{\partial \eta^2}\big)^{-1}$ where

$$\frac{\partial L}{\partial \eta} = kV\big(\Psi(V\eta) - \Psi(\eta)\big) + \textstyle\sum_{i=1}^{k}\sum_{j=1}^{V}\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{l=1}^{V}\gamma_{il})\Big), \qquad \frac{\partial^2 L}{\partial \eta^2} = kV\big(V\Psi'(V\eta) - \Psi'(\eta)\big)$$

In this setting we do not calculate β, but we can estimate its expectation using λ, $\mathbb{E}_q[\beta_{ij} \mid \lambda] = \lambda_{ij}/\sum_{l=1}^{V}\lambda_{il}$.

1.4 Extension of LDA - "Topics over time"

As Blei et al. mention in the concluding remarks of the original article, one of the main advantages of the generative model is it's modularity: LDA can be a module in a more complex model (LDA paper). Indeed there have been numerous extensions to the basic LDA in this way since the release of the original paper in 2003 [6]. One notable extension is the author-topic model (2004) in which LDA is extended to include authorship in two ways: that an authors tend to write about certain topics and that multiple authors might have written a single document [9]. Another extension of the model that is especially interesting for our work in this article is that of topic evolution: visualizing how the the prevalence of topics in a corpus changes over time. The idea of a dynamic topic model - explicitly modelling (2006) θ and β as drawn from probability distributions changing/evolving over time - has been explored by one of the authors of the original paper, where a corpora evolving over 100+ years was studied [2]. However, on shorter time-scales ($\sim$10 years) post hoc examination of how the θ parameters evolve over time have shown good results qualitatively in a static model [4]. Other improvements have of course also been made and today LDA is one of the most popular methods for topic modelling [6].

1.5 Scope and objectives

In this report our main aim is to reproduce the results in two of the experiments mentioned above: document classification and collaborative filtering. We don't reproduce the document modelling experiment since the data was unavailable to use given our project budget. Beyond this, we investigate LDA:s usability for estimating the evolution of "hidden" topics over time: how "much" is written about a certain topic and how this changes over time.

2 Method

```
Initialize α and β randomly;
while L(γ, φ; α, β) not converged do
    E-step;
    for each document d in corpus do
        Initialize γ_d randomly;
        while γ_d not converged do
            Calculate and normalize φ_dni^{t+1};
            Calculate γ_di^{t+1};
        end
    end
    M-step;
    Update α with NR;
    Calculate and normalize β;
end
```

Algorithm 1: Pseudo-code for un-smoothed LDA

```
Initialize α, η and λ randomly;
while L(γ, φ, λ; α, η) not converged do
    E-step;
    for each document d in corpus do
        Initialize γ_d randomly;
        while γ_d not converged do
            Calculate and normalize φ_dni^{t+1};
            Calculate γ_di^{t+1};
        end
    end
    M-step;
    Calculate λ_ij;
    Update α and η with NR;
end
```

Algorithm 2: Pseudo-code for smoothed LDA

The unsmoothed and smoothed model was implemented in python using the update equations defined earlier and according to algorithm 1 and 2 respectively. Convergence in γ_d was defined as $\frac{1}{k}\|\Delta\gamma_d\| < 0.00001$. The same criteria was used for the Newton Raphson updates, however, convergence in the lower bound was defined as the relative improvement being < 0.0001. Since every iteration of the VI is $\mathcal{O}((N_d+1)k)$ (in practice around N^2k operations to convergence per document)[3], we clearly see that the algorithm becomes slow for large corpora with long documents. We made it more efficient by using matrix multiplication, which made the unsmoothed version fast enough to run on all (complete) datasets by making the convergence criteria more lenient (0.000001 in the original paper). However, the smoothed version, which has an additional variable to converge, was only fast enough to be run on the smaller EachMovie dataset.

3 Applications and Results

3.1 Latent word topics in an article

One way of getting insights in an article the LDA model has trained on is looking at its variational parameters γ and ϕ. Looking at the update equation for γ we see that $\gamma_i - \alpha_i$ indicates the expected number of words belonging to topic i in that article. Thus, topics for which $\gamma_i - \alpha_i \geq 1$ can be seen as "main" topics of that article. Furthermore, looking at large values of ϕ_{ni} (> 0.9) we find words that are very likely to belong to topic i since ϕ_n approximates $p(z_n \mid \mathbf{w})$. This was done for an article in the Guardian corpus (more on this in 3.4) with $k = 150$, seen in figure 5a, where the words are colored by their latent topics.

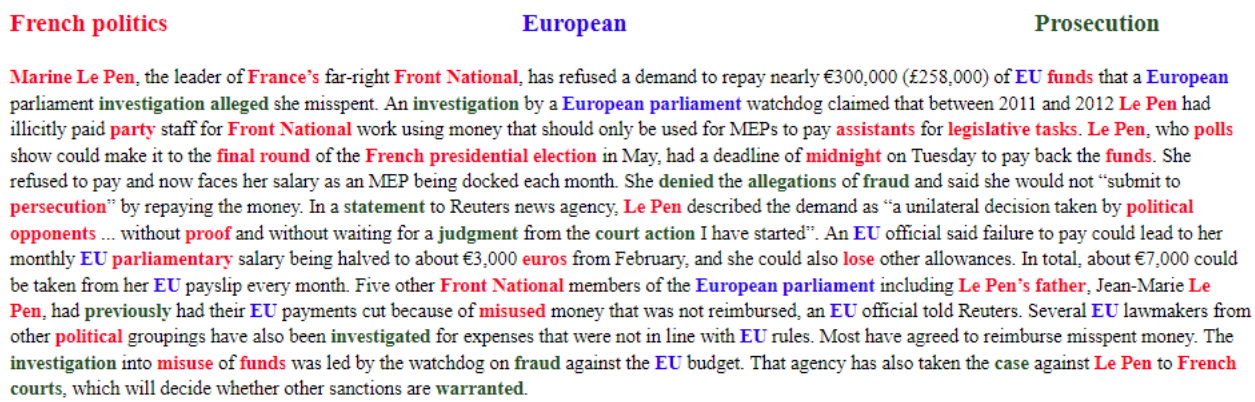


Figure 3: An example article from the Guardian corpus. Each color indicates from which latent topic the word was putatively generated.

3.2 Document Classification

The second experiment we wanted to replicate was that of binary document classification using the Reuters-21578 corpus. It was a bit unclear which articles in the corpus had been used in the paper's experiment: the full dataset – as the name implies – has 21578 articles from Reuters' newswire in 1987 but the authors write that they only used 8000 documents containing 15818 unique words. However none of the recommended selections in the documentation of the dataset yielded those numbers. In order to get a document selection that is both similar to the author's and easily reproduceable for future work we chose to work with "The Modified Apte" (ModApte) selection of articles. More specifically we selected all articles in the Reuters set belonging to the ModApte training set that had a topic assigned to them (a label). This data set had 7183 articles and 23498 unique words.

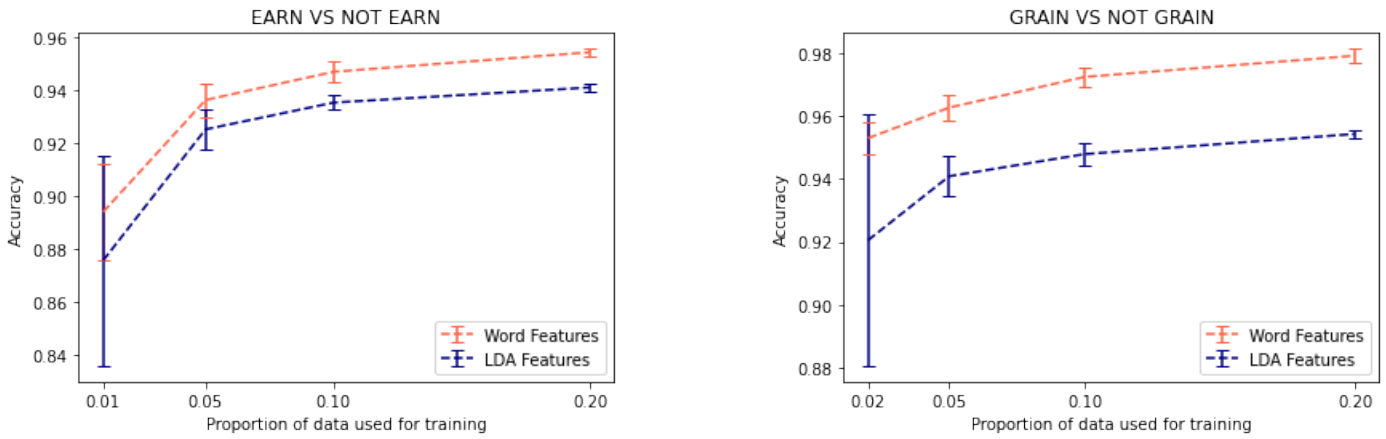
Without a clear description of how the authors had done text preprocessing, we thought it be best to use preprocessing conservatively: each article was tokenized with Natural Lanaguage Toolkit (NLTK) using nltk.sent_tokenize() and nltk.word_tokenize(), stopwords removed using nltk.stopwords, tokens/words with only non-alphanumerical characters were removed and finally we normalized all tokens/words with numbers in them to a single number token. We did this with the aim of decreasing "noise" in the dataset so that the LDA algorithm could better discriminate between topics. No lemmatizer or stemmer were used, and no named entity recognition was done.

After running the unsmoothed LDA with k=50 on this dataset we compared an SVM trained on the resulting Dirichlet parameters $\gamma*$ to a SVM trained on all the word features of an article (the counts of all words in the vocabulary, in the document) for the two classification tasks. For the Dirichlet features we used grid-search and cross-validation to train the SVM (sklearn implementation) with the three basic kernels (linear, polynomial and gaussian) whereas for the word features we tuned only a linear SVM due to running time issues with the increased number of features. In each of the four graphs the hyperparameters yielding the highest accuracy at training set proportion 0.2 were selected.

As can be seen in figure 4 our results – albeit not identical – are similar to the ones in the paper. Indeed, we find that much discriminatory information is kept when LDA is used to reduce the dimensionality of a document, from 23498 to 50 features. Since the author's method description leaves a lot to be desired, it's not surprising that the results deviate slightly. Beyond the fact that we used different parts of the same dataset it is also unclear what kind of text preprocessing had been done, if they used custom kernels, how they tuned the algorithms and what evaluation metric they used. The first two are maybe especially important since they affected the input to the algorithm, which could have had a large impact specifically on the Grain vs No Grain where the classes were very unbalanced.

3.3 Collaborative Filtering

Our third experiment was on the EachMovie collaborative filtering data where users indicate their preferred movie choices. Here a user can be seen as a document and their preferred movies as words. The task becomes to predict the



(a) (b)

Figure 4: Results for the two classification problems. In (a) the best kernel was a Gaussian with C=1000, γ=0.001. In (b) the kernel with the highest accuracy was also a Gaussian with C=1000, γ=0.0001

probability of a held-out movie of a set of test users when we know the rest of their preferred movies (note that this can be interpreted as a recommendation task). Before this the model has been trained on a training set with no held-out movies. To evaluate the models accuracies we use the predictive perplexity defined as follows:

$$\textit{predictive-perplexity}(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log(w_{d,N_d} \mid \mathbf{w}_{d,1:N_d-1})}{M} \right\}$$

We trained a mixture of unigrams model, the unsmoothed LDA and the smoothed LDA on this data. For these models we get the following probabilities of a movie given a set of observed movies:

$$p(w \mid \mathbf{w}_{obs}) = \sum_z p(w \mid z)p(z \mid \mathbf{w}_{obs}) \qquad p(w \mid \mathbf{w}_{obs}) = \int \sum_z p(w \mid z)p(z \mid \theta)p(\theta \mid \mathbf{w}_{obs})d\theta$$

respectively. Using that $p(w \mid z) := \beta_{zw}$, $p(z \mid \theta) \sim \text{Multinomial}(\theta)$ and that $p(\theta \mid \mathbf{w}_{obs})$ can be approximated by $q(\theta \mid \gamma) \sim \text{Dir}(\gamma)$ from the variational algorithm when it's re-run on $\mathbf{w}_{obs}$ we can simplify the expression. By interchanging the sum and integral sign we get

$$p(w \mid \mathbf{w}_{obs}) \approx \int \sum_z \beta_{zw}\theta_z q(\theta \mid \gamma)d\theta = \sum_z \beta_{zw} \int \theta_z q(\theta \mid \gamma)d\theta = \sum_z \beta_{zw}\, \mathbb{E}_q\left[\theta_z\right] = \sum_z \beta_{zw} \frac{\gamma_z}{\sum_{i=1}^{k} \gamma_i} \tag{7}$$

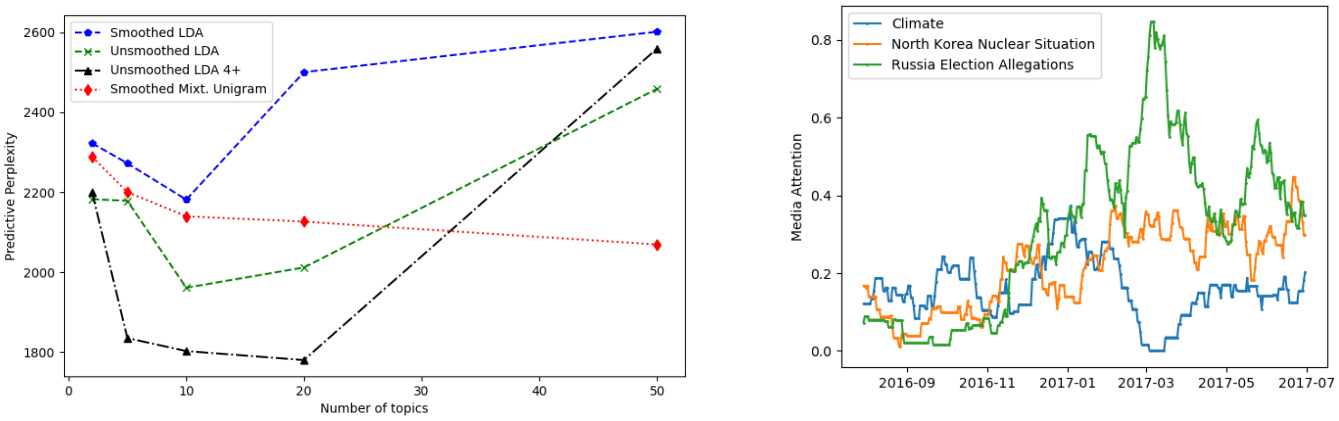
that is a linear combination of k Dirichlet expectations. In the smoothed setting β_{zw} is exchanged with $\mathbb{E}_q[\beta_{zw}] = \lambda_{zw}/\sum_{l=1}^{V} \lambda_{zl}$.

The results can be seen in figure 5a. In the original report they specify that they restricted the dataset to users that had rated at least 100 movies with at least four out of five stars. Following this criteria we only get 1330 users. Setting the criteria to at least three out of five stars instead we get 3680 users, which is in line with the paper. Thus, we trained the models on both these datasets, where 4+ denotes the first, to make the results comparable.

3.4 Topics over time

To conclude our acquaintance with the LDA model we were interested in using LDA to discover topics from a corpus and visualize how the prevalence of those topics in the corpus changed over time. As this is an unsupervised learning problem – there are no labels indicating how "correct" an estimated topic progression is – we wanted to have some sort of qualitative indication of how good the results are. For this reason we choose to work with newspaper articles where we found the "All The News" dataset on Kaggle[7], containing articles scraped from the Guardian's main webpage from July 2016 to June 2017. We preprocessed the data in the same way as in the Classification experiment. The second advantage of working with newspaper articles for this task is that they cover diverse things such as sports, politics, climate, something that was a problem with the Reuters corpus where both articles and the estimated word distributions β_i where very homogenous. After training the unsmoothed LDA model (k=120) we used the θ-vectors of articles in a time period to measure the topic prevalence/media attention given to the k different topics in the timeframe using the following estimation:

$$\theta^*_{m_1} + \theta^*_{m_2} + \cdots + \theta^*_{m_s} \approx \sum_{d \in [M]} \mathbb{E}_q[\theta_d]$$



(a) (b)

Figure 5: **Left:** Predictive perplexity for the EachMovie dataset for different models. 4+ represents the smaller dataset. **Right:** Estimated topic prevalence (here called Media Attention) of three interesting topics from June 2016 - July 2017 in the Guardian Corpus

Here $[M] = \{m_1, ..., m_s\}$ denotes the indices of all articles in a given timeframe, $\theta^*_{m_1}$ denotes the "actual" hidden value of θ - the topic proportions of article m_1 - and q is our free variational Dirchlet distribution over *theta*. Using this for each day, week, month (or other timeframe) we get an estimate of how the prevalence of all k topics in the corpus changes over time. Using this formula on a daily basis with a 30-day rolling average to get smoother results and choosing three interesting topics we get figure 5b.

To evaluate the validity of our results we have identified major news stories which coincide with the peaks of the topic prevalence over time in figure 5b. Regarding climate, the peak is subjected around January 2017, which was a period when the Trump administration actively worked to erase ex-president Barack Obama's climate initiatives[11]. The graph of the Russian Election Allegations which gained large media traction after members of the United States Congress publicly disclosed a potential russian interference in the 2016 election. The media attention continued to increase as more US intelligence agencies confirmed the suspicions and the Office of the Director of National Intelligence published its detailed report in January 2017[8]. Notably, the last peak from 2017-05 to 2017-07 coincides with the news surrounding Trump's dismissal of James Comey, the FBI director, in may 9th, 2017[10].

4 Discussion

Overall, reproducing the LDA experiments in [3] has been a challenging task, as the specification of the data pre-processing of the original article is ambiguous, both for the Reuters' dataset and the EachMovie users, making exact reproduction difficult. Nevertheless, similar conclusions could be reached in some situations from our performed experiments, demonstrating that LDA could outperform other generative models in some tasks.

However, it is shown in Figure 3, that the LDA bag-of-words model assumption indeed allows to distribute different words from the same concept into several topics, as it occurs with *EU* and *parliamentary* in figure 5a. This fact not only misses the real meaning of some composed syntagmas, but also could hinder classification or the distribution of topics over the time, by enforcing topics that are not actually present in a document. As it is proposed on the paper, maintaining some kind of possible sequentiality between consecutive words, maybe based on n-gram Markov assumption (the probability of a word depends on the n previous words), could solve this issue. Nevertheless, it could at the same time further increase the complexity of LDA model, making it difficult to derive its update equations for all the parameters.

Another problematic aspect of the standard LDA version is its huge computational complexity. In each EM iteration, all the variational variables are re-estimated using the VI algorithm, which makes the complete algorithm $O(M \times N^2 \times k)$, M being the size of the corpus and N the maximum length of the documents, and k the number of topics. In practice, for large datasets with long documents the algorithm took several hours to converge, even with the leneient convergence criterion. This could be a problem when more documents are added to the corpus, which could help the model being updated when it is trained with journalistic texts.

The LDA models presented in this report are not only computationally heavy, they are also inflexible regarding adapting to new data. As the models have been defined, their parameters need to be re-fitted for any addition of new data, making them ill-suited for data streams. A later paper, in which Blei once again is one of the authors, developed an online learning version of LDA. Being based on online stochastic optimization with a natural gradient step, it can easily handle larger corpora including data streams.[5]

Comparing our example article from the Guardian corpus to the example article from the AP corpus we see that our example has a lower density of highlighted words. This could possibly be explained by the pre-processing of the

data, since in this example around 40% of the total words were removed. Notably we observe that all numbers in the AP article are highlighted while none are in our example. This could be due to our handling of numbers as we gave all numbers a dummy word, regardless of context, e.g. $ or € before the number.

Continuing on the example article we would like to challenge the authors reasoning regarding the main topics of an article. As argued, $\gamma_i - \alpha_i$ gives the expected number of words belonging to topic i. Setting the criteria at $\gamma_i - \alpha_i \geq 1$, we would insist that one would capture more than just the main topics of the article. We instead suggest raising the threshold, preferable setting it to a proportion of the length of the article, since a single (or a couple) word belonging to a topic is not enough to make it a main topic. We performed small examples of this, which (not surprisingly) gave fewer and more accurate core topics.

Looking at the collaborative filtering; when we combine the ambiguous description of the dataset in Blei's paper and our results, where all models get a predictive perplexity of around 2000, we strengthen our suspicions regarding our data pre-processing differing from the original paper. However, we see that the smoothed mixture of unigrams performs better than smoothed LDA and better than the other models for large k's. This is not in line with the original paper, and can probably be caused by our lenient convergence criteria on the lower bound. Furthermore, we see that LDA on 4+ performs the best, which is not surprising since narrowing the definition of 'preferred' movies should improve predictions. We do however see a similar pattern as the original paper, all LDA models seem to perform best for $k = 10, 20$. Also note that the predictive perplexity should explode for the unsmoothed LDA when there are documents/movies in the testset that the model hasn't trained on. For this case, when the vocabulary is really small ($V = 1600$) there are only two such occurrences, which we solved heuristically by assigning unseen movies probability $\frac{1}{V}$ post fitting.

Regarding the analysis of media attention given to different topics over time it's interesting to note that the media attention estimated ad hoc with the static LDA model seems to be decently backed by the news events and media reporting that took place: atleast in terms of relative media attention. We did not have to incorporate change in our model compared to the dynamic topic model [2] which modelled change in both the distribution over β and over θ. One possible explanation lies in the timeframe: since the articles are from a single year and not 100+ years a modelling change might not be as necessary. Another potential explanation is that media reporting around a topic often rapidly changing in spurs rather than smoothly and systematically over time, compared to the science articles from the past 100+ years studied in [2].

References

[1] David Blei. Topic models. Lecture, Cambridge University, Summer School, 2009. `http://videolectures.net/mlss09uk_blei_tm/`.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[5] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 856–864. Curran Associates, Inc., 2010.

[6] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[7] Kaggle. All the news, 2017. `https://www.kaggle.com/snapcrack/all-the-news?select=articles1.csv`.

[8] Office of the Director of National Intelligence. Assessing russian activities and intentions in recent us elections. Intelligence Community Assessment, 2017. `https://www.dni.gov/files/documents/ICA_2017_01.pdf`.

[9] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*, 2012.

[10] Michael D. Shear and Matt Apuzzo. F.b.i director james comey is fired by trump, 2017. `https://www.nytimes.com/2017/05/09/us/politics/james-comey-fired-fbi.html`.

[11] Valerie Volcovici. Trump administration tells epa to cut climate page from website: sources, 2017. `https://www.reuters.com/article/us-usa-trumpepa-climatechange/trump-administration-tells-epa-to-cut-climate-page-fromwebsite-sources-idUSKBN15906G`.