Topic models for short text data

Silviu Paun

A thesis submitted for the degree of

Doctor of Philosophy

School of Computer Science and Electronic Engineering

University of Essex

February 2017

Abstract

Topic models are known to suffer from sparsity when applied to short text data. The problem is caused by a reduced number of observations available for a reliable inference (i.e.: the words in a document).

A popular heuristic utilized to overcome this problem is to perform before training some form of document aggregation by context (e.g.: author, hashtag). We dedicated one part of this dissertation to modeling explicitly the implicit assumptions of the document aggregation heuristic and applying it to two well known model architectures: a mixture and an admixture. Our findings indicate that an admixture model benefits more from aggregation compared to a mixture model which rarely improved over its baseline (the standard mixture). We also find that the state of the art in short text data can be surpassed as long as every context is shared by a small number of documents.

In the second part of the dissertation we develop a more general purpose topic model which can also be used when contextual information is not available. The proposed model is formulated around the observation that in normal text data, a classic topic model like an admixture works well because patterns of word co-occurrences arise across the documents. However, the possibility of such patterns to arise in a short text dataset is reduced. The model assumes every document is a bag of word co-occurrences, where each co-occurrence belongs to a latent topic. The documents are enhanced a priori with related co-occurrences from the other documents, such that the collection will have a greater chance of exhibiting word patterns. The proposed model performs well managing to surpass the state of the art and popular topic model baselines.

Acknowledgements

I am grateful to my supervisors, Udo Kruschwitz and Massimo Poesio, for letting me pursue my own academic path and for supporting me along the way. Their tolerance in the face of my excitement for sometimes divergent theoretical aspects is something I am thankful for as it offered me the necessary freedom to develop.

This dissertation has been funded by an EPSRC CASE award studentship with BT plc, whose financial and technical support I gratefully acknowledge. I am thankful to Paul Mckee and Mike Fisher, who just like my supervisors, did not impose any constraints to my academic path. Special thanks to John Davies as well for his support in the early stages of my career - I spent a wonderful summer internship at BT!

I would also like to thank my parents, Anelize and Gheorghe Paun, for their love and support along the years. I am grateful as well to my grandparents, Cernovica and Nicolae Paun.

To my grandparents, Ana and Ion Paun.

Contents

Chapter 1

Introduction

A great number of text collections are already available or being produced with high velocity and in large volumes, having the potential to offer value to people. Examples of such collections include digital libraries of scientific publications, news articles, books, blogs, web pages or social media posts. However, extracting useful information from large unstructured datasets remains challenging and automatic methods for doing so are essential. Topic models are a very promising way of structuring the data in an automatic fashion to make it available to end users in a more easily digestible format.

A topic model, at its core, is a probabilistic method for extracting the main themes from an unstructured collection of text. It offers end users the opportunity to search and explore data in ways beyond the traditional keyword-based queries. For example, a digital library may contain millions of documents from heterogeneous topics such as literature, biology or mathematics. A topic model could automatically detect the existence of such themes, and much more. It can allow a user to focus only on the documents part of the literature theme. It can go further and identify finer grained topics of this theme like fiction, comedy or drama. In a standard framework (e.g.: a mixed membership model like Latent Dirichlet Allocation [9]) a topic model offers two kinds of information: 1) the identified topics represented by a probability distribution over the vocabulary space where the descriptive words are those with high probabilities; and 2) the coverage of each topic in the documents.

Topic models have been shown to have wide applicability [5, 6, 8]. A few examples

include analyzing the evolution of topics over time in digital library data [7], the identification of correlated topics [4, 25], modeling authors and their publications [40] or capturing spatial and temporal patterns from blog posts [33]. Topic models can also be utilized to get a low dimensional semantic representation of the documents. This can be useful in document clustering or classification tasks [27].

In this dissertation we are concerned with topic models for short text data, an emergent area of research [26, 39, 56, 61]. This type of data where text items are short compared to traditional documents like a published paper or a news article, is present in many environments; examples include tweets, titles of scientific publications, of blogs, of news, forum conversations or short product reviews. In this text environment traditional topic models like LDA under-perform. The problem is caused by a reduced number of observations available for a reliable inference (i.e.: the words in a document). This causes topic models to suffer from sparsity.

Researchers have addressed sparsity from multiple angles: 1) context has been leveraged to aggregate documents before training the models [22, 31]; 2) general purpose models have been built which can be used when contextual information is not available [39, 56, 61]; and 3) the short text documents are enhanced a priori to the learning phase with external information [49, 50]. In this dissertation we touch on all three aspects: we introduce two models that account for context; and we build a general purpose model which enhances the documents with extra information, but the information is generated internally, from the input collection. The remainder of this chapter is dedicated to guiding the thesis.

We cover the related literature in Chapter 2. We begin with some basic theory about a mixture and an admixture model. This is followed by a review of a broad range of topic models to showcase their wide applicability. We then discuss models developed for short text data, the research focus of this dissertation. The final parts of the chapter cover a review of parameter estimation techniques, with a focus towards variational inference, the technique employed in this thesis. We conclude the chapter with a discussion on common evaluation methods.

After introducing the chapter on related literature, we now formulate the research questions which stand at the core of this dissertation. Each question is introduced

below, starting from appropriate observations, and is followed by the chapter in which it is addressed.

It is known that document aggregation by context helps LDA (the admixture) to alleviate sparsity in short text data [22, 31]. At the same time, Mixture of Unigrams (the mixture) has become a popular baseline in this area [26, 36, 56, 61]; its one topic per document assumption making it attractive for short text items [59]. More than that, the mixture and the admixture are standard classes of models found at the core of a wide variety of topic models developed over the years (Chapter 2 covers an in-depth review). With these observations in mind, we formulate the following research questions:

Which class of models benefits more from aggregation in short text data, a mixture or an admixture? Can document aggregation lead to state of the art performance?

We address these questions in Chapter 3 where we explicitly model the implicit assumptions of document aggregation, and apply it to the two standard model architectures. We evaluate the enhanced models on both very short (i.e.: titles of publications) and medium (i.e.: abstracts) text items, with different opportunities for aggregation (a smaller vs. a larger number of documents per context). The evaluation targets multiple tasks such as topic coherence, document clustering and document classification. Our findings indicate that an admixture model benefits more from aggregation compared to a mixture model which rarely improved over its baseline (i.e.: the standard mixture). We also find that the state of the art in short text data can be surpassed as long as every context contains a small number of documents.

Contextual information is not always available or it does not help (i.e.: it is shared by documents which have little or no topical relationship). In these cases, a general purpose topic model is desirable. In normal text data, a classic model like LDA works well because patterns of word co-occurrences arise across the documents. However, the possibility of such patterns to arise in a short text dataset is reduced. Based on this observation we formulate the following research question:

Can short text collections be enhanced such that repeating word co-occurrences have a better chance to arise across the documents more consistently and facilitate

a better topic discovery?

We address this question in Chapter 4 where we introduce a new topic model for short text items. The model assumes every document is a bag of word co-occurrences, where each co-occurrence belongs to a latent topic. The documents are enhanced a priori with related co-occurrences from the other documents, such that the collection will have a greater chance of exhibiting word patterns. We evaluate the model on two labeled datasets of tweets and one of titles of scientific publications. The latter is a dataset which we also utilized in Chapter 3 and has contextual information available. We target in the evaluation multiple tasks such as topic coherence, document clustering and document classification. The model we propose performs well managing to surpass the state of the art and popular topic model baselines. The best performing contextual model introduced in Chapter 3 managed to get the best results in this evaluation as well, further strengthening the argument that contextual information is indeed useful when available.

In the previous chapters, the approaches taken to alleviate sparsity were oriented towards increasing the number of observations (i.e.: the words) available for the inference of the K-dimensional vectors governing the topic proportions (where K is the total number of topics). Considering these vectors are known to be the main reason behind LDA's poor performance in short text data (point also raised by Yan *et al.* [56]), a different approach is worth investigating:

Can topic models be improved by assuming a more appropriate number of topics for every document?

We address this question in Chapter 5, where we experiment with a topic model which assumes documents are mixtures of only a subset of the entire topic space. This complements existing work which assumes documents contain either a single topic or a mixture of the entire topic space. The main motivation behind this chapter is that neither of the aforementioned assumptions are entirely plausible. Even if the "one topic per document" assumption performs reasonably well on a short text dataset such as a Twitter collection, there can be many tweets which cover more than one topic. At the same time, even though longer documents tend to cover multiple topics, it is implausible they cover the whole topic space. The evaluation

assesses coherence, a measure of topic interpretability, and is performed in varying text environments from very short to medium and longer text. The experiments indicate a connection between the size of the documents and the performance of the models with respect to the number of topics assumed for every document.

We conclude the dissertation with Chapter 6. A summary of the key points is given reiterating the novelty brought by this work and its applicability.

Chapter 2

Related Work

2.1 Topic Models

Text data is being generated at a rapid pace and is available in a variety of environments such as social platforms, digital libraries or the media. For end users to be able to interact with the large amounts of available data, some form of data organization is needed. Topic models offer one way of structuring the data in an automatic fashion to make it available to people in a more easily digestible format. The organization is done based on themes identified at different levels of granularity. Such generative models of text have been developed over the years for a wide variety of applications [5, 6, 8, 25, 32, 41]. The literature on topic models is so extensive that only a partial discussion is possible.

We begin with some basic theory about topic models. This is followed by a review of a broad range of models to showcase their wide applicability. We then discuss topic models for short text data, the research focus of this dissertation. The final parts of the chapter cover a review of parameter estimation techniques, followed by a discussion on common evaluation methods employed in the literature.

2.1.1 Core Models

In this section we describe two core models used for discovering latent topics from text collections: a mixture and an admixture model. The latter is also known as a mixed membership model. The difference between the two models is that in an

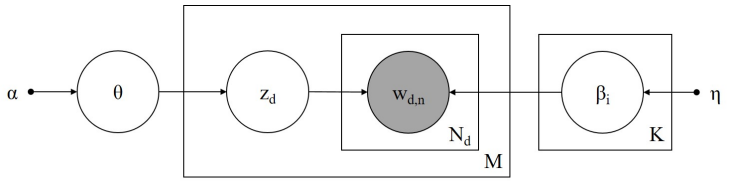


Figure 2.1: Graphical Model of MoU

admixture a document can exhibit multiple topics, whereas in a mixture documents are assumed to be generated from only one topic. These basic architectures have laid the foundation for a variety of models over the years (see Sections 2.1.2, 2.1.3 and 2.1.4 for an in-depth review).

The models take as input a collection of documents indexed by $d \in \{1, 2, ..., M\}$. Every document d is a collection of words indexed by $n \in \{1, 2, ..., N_d\}$. The words form a vocabulary space indexed by $j \in \{1, 2, ..., V\}$.

2.1.1.1 The Mixture Model

In this section we describe Mixture of Unigrams (MoU) [37], a basic but popular model for latent topic identification. The model is known for its "one topic per document" assumption which makes it a strong baseline in short text data [36, 56].

The graphical model of MoU is presented in Figure 2.1. The generative process is given below:

- For every topic $i \in \{1, 2, ..., K\}$:
  - Draw a word distribution $\beta_i \sim Dir(\eta)$
- Draw global topic proportions $\theta \sim Dir(\alpha)$
- For every document $d \in \{1, 2, ..., M\}$:
  - Draw a topic $z_d \sim Cat(\theta)$
  - For every word position $n \in \{1, 2, ..., N_d\}$:
    - Draw word $w_{d,n} \sim Cat(\beta_{z_d})$

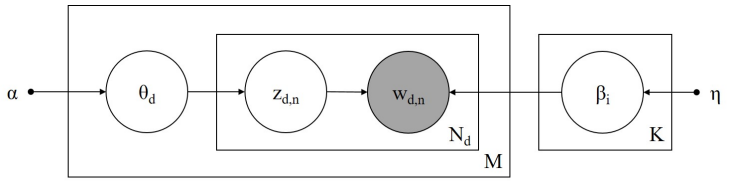


Figure 2.2: Graphical Model of LDA

2.1.1.2 The Admixture Model

In this section we describe Latent Dirichlet Allocation (LDA) [9], a well known topic model where a mixture of topics is responsible for generating the words in a document. LDA can be considered a more general Bayesian extension of the Probabilistic Latent Semantic Analysis (PLSA) model published previously by Hofmann [21]. In PLSA, there is no assumption that guides the generative process of the document specific topic proportions. Blei *et al.* [9] makes the observation that PLSA is unsuitable for prediction tasks on unseen documents and that it is prone to overfitting.

LDA has become the backbone of a wide variety of topic models over the years (see Section 2.1.2). Supporting material can be found in numerous previous studies [8, 16, 44]. Compared to MoU, LDA relaxes the one topic per document assumption.

The graphical model of LDA is presented in Figure 2.2. The generative process is given below:

- For every topic $i \in \{1, 2, ..., K\}$:
  - Draw a word distribution $\beta_i \sim Dir(\eta)$
- For every document $d \in \{1, 2, ..., M\}$:
  - Draw document-level topic proportions $\theta_d \sim Dir(\alpha)$
  - For every word position $n \in \{1, 2, ..., N_d\}$:
    - Draw a topic $z_{d,n} \sim Cat(\theta_d)$
    - Draw word $w_{d,n} \sim Cat(\beta_{z_{d,n}})$

2.1.2 Topic Models: A Broad Survey of the Literature

The wide applicability of topic models is known and has been reviewed extensively in previous work [5, 6, 8]. Nevertheless, for completeness, we will discuss a selection

of papers to show how topic models have been adapted over the years.

The set of assumptions that form the core of an admixutre model like LDA have been adapted in various ways as richer models have been developed. One such example is the Bigram Topic Model proposed by Wallach [51], where documents are not viewed any more as simple bags of words: a bigram language model now guides the generative process of the words given the topics. Another example is the Dynamic Topic Model proposed by Blei & Lafferty [7] where the order of the documents in the collection is taken into account (this is in contrast with LDA where the order does not matter). The model aims to analyze the evolution of topics over time in a large collection of documents. In the generative process, the topics are assumed to evolve from one time slice to another with Gaussian noise. Hierarchical priors shared by the topic proportions of the documents part of the same time slice, evolve as well with Gaussian noise. A further example of modeling outside the standard assumptions of LDA, is to allow the complexity of the data to determine the number of topics in a collection. The Hierarchical Dirichlet Process is one such model example [20, 53].

Researchers have also focused on modeling potential correlations between topics, another limitation of LDA. For example, Blei & Lafferty [4] model document specific topic proportions with the help of a logistic normal distribution. The covariance matrix of the just mentioned distribution is responsible for capturing the correlations. An alternative to this model is the one developed by Li & McCallum [25] where correlations can be captured with an arbitrary directed acyclic graph (e.g.: structures where super topics have correlated sub-topics).

Topic models have also been developed to capture patterns beyond simple word co-occurrences. For example, Wang & McCallum [54] proposed a generative model which learns, in addition to the topics, beta distributions that capture their trends over time. The model assumes that for every word position in a document a topic is drawn, then the word is drawn from that topic, followed by the timestamp of the document. This way, the topics from the documents are influenced by both the words and their timestamps. In a different context, models have also been developed to account for both short range syntactic and long range semantic dependencies [17].

Documents are broken down into function and content words. The function words are captured by a Hidden Markov Model while the content words are handled with the help of a topic model, all in a unified generative model that integrates topics with syntax.

With the increasing popularity of word embeddings researchers have started exploiting them in topic models as well. Nguyen *et al.* [36] propose two extensions of the popular topic models MoU and LDA. The extensions include in the standard models a latent feature component. The generative process assumes the words are being draw either from the classic topic distributions or from this newly added latent feature component. The component is a categorical distribution where the probability of a word is proportional to the dot product between its embedding and a latent vector representation of its assigned topic. The authors use word embeddings pre-trained on large external corpora. Their findings indicate that the proposed models have increased performance especially on small datasets or datasets that consist of short text documents. Another model which utilizes word embeddings is GaussianLDA [14]. The model replaces the categorical distributions used in LDA to represent the topics with multivariate Gaussian distributions defined over the embedding space. This particular choice of representing the topics is a way of suggesting to the model to assign words that have similar embeddings (i.e.: vector representations; spatial similarity) to the same topic.

2.1.3 Topic Models for Context-Accompanied Text Data

Context has been extensively exploited in topic models for text mining purposes. Zhai *et al.* [58] developed a model for cross collection topical analysis. The model identifies collection-specific topics but also general topics which arise across the datasets. In another contextual model proposed by Mei *et al.* [33] the generative process assumes time and location specific topic proportions which guide the per word topic assignments of the documents. This allows the model to capture topical trends with respect to time and location. Another research effort focuses on defining a more general purpose topic model for contextual text mining [32]. The model assumes context specific topic proportions and context specific views of the topics.

The generative process also assumes that multiple contexts can be responsible for selecting the per-word topic assignments. The described models [32, 33, 58] are built as extensions of the popular PLSA [21] baseline.

The author topic models are another class of contextual models developed over the years [29, 30, 40, 41]. Every document is accompanied by an observed set of authors. The model proposed by McCallum [30] assumes that documents are generated from the word distributions that correspond to their authors (i.e.: one author per word; word drawn from author-specific word distribution). Rosen-Zvi *et al.* [40] proposed another model which takes into account word distributions associated with the topics and topic proportions associated with the authors. For every word, an author is selected, followed by a topic assigned based on the proportions that correspond to the previously selected author; with the topic at hand, the word is drawn afterwards from the appropriate topic distribution. McCallum *et al.* [29] extends the author topic model of Rosen-Zvi *et al.* [41] by incorporating recipients. The model is useful for an analysis of email data, for example. It assumes every document contains an observed author and multiple recipients. For every word, a recipient is first selected; then a topic is being drawn according to proportions that correspond to the author of the message and the assigned recipient; finally the word is drawn from the appropriate topic distribution. The described author models [29, 30, 40, 41] are built as extensions of the popular LDA [9] baseline.

2.1.4 Topic Models for Short Text Data

Probabilistic topic models for short text data are the research focus of this dissertation. The poor performance of standard models like Latent Dirichlet Allocation (LDA) on short text items is caused by sparsity [24, 36, 56, 61]. Because of the reduced number of observations per documents (i.e.: the words) the inference of the K-dimensional vectors governing the document-specific topic proportions can be unreliable. In a study on the factors which affect the performance of LDA, Tang *et al.* [45] conclude that poor performance is expected when the documents are too short, even if you have a large collection.

One popular heuristic employed by researchers to overcome sparsity in short text

data is to utilize various message aggregation strategies before training LDA. Hong & Davison [22] find that aggregating tweets based on author gives better performance over standard LDA. In a later published paper, Mehrotra *et al.* [31] found that aggregating by hashtag brings even more benefits to LDA's performance. Various other researchers who do not study the benefits brought by aggregation to their models but want to avoid sparsity in short text data employ this heuristic [49, 50, 55]. We also want to highlight here TwitterLDA [59], another frequently cited model on aggregation, which combines a mixture model with user-specific topic proportions and a background word distribution. In a more recent publication, Sasaki *et al.* [42] introduced an improved version of TwitterLDA which models user-specific preference for functional vs topical words (as opposed to the global preference from the original paper).

More general purpose topic models built for short text data also exist and can be applied when contextual information is not available. An example is Mixture of Unigrams, which models global topic proportions (unlike document-specific ones like in LDA), and is one of the first successful examples of alleviating sparsity, becoming over the years a standard baseline in this area [26, 36, 56, 61]. Its "one topic per document" assumption seems to fit reasonably well short text items. We use this model ourselves as one of the baselines in the experiments. Another popular model for short text is the Biterm Topic Model (BTM) [56]. The model has a preprocessing step in which all the biterms (i.e.: word pairs) of every document are generated. The biterms become then the input of a Mixture of Unigrams model. BTM alleviates sparsity because, just like MoU, it assumes global topic proportions. Unlike MoU which assumes one topic per document, BTM is more flexible as it assumes one topic per biterm. Since documents contain multiple biterms, they can potentially exhibit more topics. We often use this model to represent the state of the art in an evaluation. BTM has been extended in various ways more recently. For example, in one extention, Yan *et al.* [57] take into account background words in addition to topical words in a model which aims to capture bursty topics from microblogs. In another example, Chen *et al.* [13] introduce Twitter-BTM, which assumes user-specific topic proportions and models as well functional words in addition to the topic

distributions. These assumptions though, make the Twitter-BTM model applicable to data where the required contextual information is available.

Going back to general purpose topic models for short text data, the Dual-Sparse Topic Model of Lin *et al.* [26] is a great example of explicitly addressing sparsity. The model keeps the usual assumptions of an admixture model, but with a twist: "Spike and Slab" priors are used to control the sparsity that may arise in both the document specific topic proportions and in the word distributions associated with the topics (hence the dual-sparse terminology). The mathematics involve constructing the Dirichlet distributions that model the just mentioned aspects in a way which allows to control which components can receive probability mass. This is achieved with the help of random Bernoulli indicators. Under this structure, the model can enforce only a few words and topics to end up with most of the mass (the others having negligible quantities). In a recent publication, Zuo *et al.* [61] propose a new general purpose topic model for short text (Pseudo-document Topic Model) which alleviates sparsity by modeling topic proportions specific to latent clusters of documents. In the generative process, for every short document, you first select a cluster and then its words are being generated according to the cluster's topic proportions. The model reduces sparsity since the topic proportions are now associated with each cluster of documents instead of having one such vector for every short text item. In the same paper, Zuo *et al.* [61] introduce another model which applies a "Spike and Slab" prior to the cluster-specific topic proportions of PTM for an explicit control of sparsity. Another model of latent document aggregation is the one proposed by Quan *et al.* [39], but there are a few differences compared to PTM. The model has a generative process which can be described as a two part mechanism: one in which large latent documents are generated from an admixture model; and a second part in which every observed short document is assumed to be generated from a latent large document.

There are also models developed to capture richer patterns that go beyond simple word co-occurrences. One such model is the Latent Event Model (LEM) proposed by Zhou *et al.* [60]. LEM models an event with multiple distributions accounting for non-location named entities, locations, time, and other descriptive words. To

alleviate sparsity, the model, similar to MoU, assumes global event proportions and one event per document. Similar models have been published for multi-faceted topic discovery in Twitter [49, 50] - in this work though the authors assume document specific topic proportions and alleviate sparsity by enhancing the short documents with external information gathered from the URLs inside the posts. They also take into consideration internal information and enhance the documents with high frequency words that appear across the collection in the company of the same hashtags [50]. Li *et al.* [24] published a new model recently which takes into account the available structure of conversations in microblogs. They first use a leader detection model to classify documents into leaders and followers. This information is used as prior knowledge to a probabilistic topic model. The generative process of the model assumes, for each message, first deciding whether it is a leader or a follower (informed by the just described prior information). If the message is a leader, a topic is drawn according to the proportions that correspond to leaders; otherwise, a topic is drawn according to proportions that correspond to the topic of the follower's leader. This separation of topic proportions indicates that leaders generate new topics and followers generate correlations between these topics. With a topic assigned to a message, its words are now generated either from a background word distribution or from the appropriate topic distribution. Leaders and followers also have their own preference for functional vs topical words. Exploiting the structure of the conversations brings improvements over other competing models (e.g.: [39, 56]) in terms of topic coherence.

2.2 Inference in Topic Models

The posterior of a topic model is often intractable for exact inference. Both deterministic and non-deterministic methods can be followed to obtain a posterior approximation.

Non-deterministic approaches include sampling techniques such as those from the Markov Chain Monte Carlo (MCMC) family. Gibbs sampling (e.g.: standard, blocked, collapsed) is one type of MCMC which is highly utilized in related work

[36, 56]. A simple Gibbs Sampler involves getting samples in an iterative procedure from the complete conditionals. The samples are then used to compute estimates of interest (e.g.: parameter means). The first few samples are usually discarded in order to move away from the random initialization to an area of high posterior density (the burn-in period). Getting representative samples is nevertheless an open problem; it is also not straightforward to assess the convergence of MCMC methods [23].

Because of these problems, in recent years, a number of researchers [7, 9, 10, 19, 20] have adopted Variational Inference, a deterministic approach to posterior approximation. This is also the parameter estimation technique employed in this dissertation. The following subsections will detail the necessary theory.

We also note here collapsed variational inference [3, 43, 46, 47]. The method is deterministic but inspired by Collapsed Gibbs Sampling - it provides a tighter bound (i.e.: the variational objective function) when compared to standard variational inference by marginalizing out some of the parameters.

2.2.1 Variational Inference

Variational Inference is a deterministic approach to posterior approximation. Let $\mathcal{M}$ be a model of some data D with parameters θ. We are going to approximate the intractable posterior $p(\theta|D)$ with a variational distribution $q(\theta)$ such that the Kullback-Leibler (KL) divergence between the two distributions is minimized.

It can be easily proved that minimizing the KL divergence between q and p is the same as maximizing the evidence lower bound (ELBO) $\mathcal{L}$ (see Equation 2.1).

$$\begin{aligned} D_{KL}(q(\theta)||p(\theta|D)) &= E_q[\log q(\theta)] - E_q[\log p(\theta|D)] \\ &= E_q[\log q(\theta)] - E_q[\log p(D,\theta)] + E_q[\log p(D)] \\ &= -\mathcal{L} + E_q[\log p(D)] \end{aligned} \tag{2.1}$$

For clarity, we express below the variational objective function:

$$\mathcal{L} = E_q[\log p(D,\theta)] - E_q[\log q(\theta)] \tag{2.2}$$

We need a variational distribution $q(\theta)$ that is tractable under expectations. In this work, we follow the common practice [9, 10, 20], and choose q to be in the mean-field variational family where each hidden variable is independent and governed by its own parameter. We review below (i.e.: Sections 2.2.1.1 and 2.2.1.2) two approaches to deriving the update formulas of the variational parameters.

2.2.1.1 Standard Variational Inference

The goal in variational inference is to maximize the objective function (i.e.: the ELBO from Equation (2.2)). The steps involved are somewhat standard for such tasks:

1. Fully expand the ELBO according to the model specification (i.e.: the appropriate full joint and variational distributions). To ease the mathematics, we represent the Dirichlet distributions in their exponential family form. It is also worth knowing that the first derivative of the log normalizer is equal to the expected value of the sufficient statistics [9].

2. With the ELBO fully expanded, the next step is to compute the update formulas of the variational parameters. The mathematics involve taking partial derivatives with respect to each parameter in question and solving the resulting equations. Note that in some cases we are dealing with constrained maximizations (e.g.: the parameters of categorical distributions) which require the usage of Lagrange multipliers.

3. With the update formulas of the variational parameters at hand, the algorithm is straightforward. The parameters are updated iteratively until the lower bound converges.

4. Monitoring the value of the ELBO is useful for assessing algorithm termination, but also for sanity checks (e.g.: the ELBO is guaranteed to increase with every iteration).

We use standard variational inference in Chapter 5. A full proof can be found in the supplemental material in Appendix C.

2.2.1.2 Variational Inference with Exponential Families

The derivations involved in obtaining the update formulas for the variational parameters can be reduced if the model in question satisfies some properties. Concretely, the model needs to have the complete conditionals in the exponential family. If this necessary condition is satisfied, getting the update formulas of the variational parameters is more straightforward because it has been proved that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals [20].

The steps involved can be summarized as follows:

1. Derive the complete conditional of every latent variable given the other latent variables and the observations. Show these are in the exponential family.

2. Define the variational distributions to have the same form as the corresponding complete conditionals.

3. Derive the update formulas of the variational parameters using the fact that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

4. With the update formulas of the variational parameters at hand, the algorithm is straightforward. The parameters are updated iteratively until convergence.

5. Monitoring the value of the ELBO is useful for assessing algorithm termination, but also for sanity checks (e.g.: the ELBO is guaranteed to increase with every iteration).

We make use of variational inference for exponential families in Chapters 3 and 4. Full proofs can be found in the supplemental material in Appendices A and B.

We note that further theory is available for this class of models. For example, Hoffman *et al.* [20] showed how to apply stochastic optimization to the variational objective function, allowing high dimensional Bayesian models to be applied at scale. Concretely, the models must have, besides complete conditionals in the exponential family, local and global parameters. In traditional batch variational inference, the

global parameters receive mass from all the local parameters (e.g.: in LDA, the "global" topics are updated using the sufficient statistics from all the "local" documents). In stochastic variational inference though, the global parameters are updated using batches of randomly selected local parameters. This way, in order to improve once the global parameters you do not have to do a full data pass. The stochastic optimization also facilitates online learning for topic models [19].

2.3 Evaluating Topic Models

The methods used to evaluate topic models differ greatly from paper to paper. Nevertheless, patterns in the choices of evaluation tasks do arise across the literature. This section reviews the tasks most utilized by researchers.

In this dissertation we evaluate the models using Topic Coherence, Document Clustering and Document Classification. These tasks enjoy wide popularity in the literature [9, 22, 25, 26, 31, 35, 56]. We dedicate a subsection for each mentioned task where we discuss in more detail the metrics used to assess it.

Besides the above mentioned methods we note there are also other ways to evaluate topic models. For example, to asses model fitness some researchers have utilized held-out perplexity [2, 9, 19]. The idea is to split the collection into train and test datasets, infer the parameters using the training data, and compute perplexity on the test set. However, computing the perplexity is intractable for topic models (because of the test set probability). To overcome this, people make use of, for example, Jensen's inequality to get a lower bound; the bound is then used as a proxy to perplexity [19]. In other cases, like in the work of Hoffman *et al.* [20], the researchers use for assessing the fitness of the model a predictive distribution in which they avoid computing such bounds. Chang *et al.* [12] found though that such methods do not correlate well with human judgment on topic interpretability. Further assessments of evaluating topic models based on the probability of held-out documents can be found in the work of Wallach *et al.* [52].

2.3.1 Topic Coherence Evaluation

One task we evaluate the models on is Topic Coherence, a measure of topic quality largely utilized in the topic models community [24, 26, 31, 56] . Newman *et al.* [35] and Mimno *et al.* [34] proposed two popular metrics utilized in the literature to measure coherence. Both metrics aim to capture the human interpretability of topics in an automatic fashion (i.e.: no human annotators). The former relies on an external corpora to compute the scores and it less correlated than the latter with human judgments [34]. The latter is also superior to word intrusion [34], another known technique to detect semantically coherent topics [12].

For the reasons explained above, we choose to utilize the topic coherence metric proposed by Mimno *et al.* [34]. Equation (2.3) lists the formula for computing the coherence score of a topic i, where $W_i = \{w_{i,1}, w_{i,2}, ...w_{i,X}\}$ is a collection of the X most probable words of that topic (in descending order) and $D()$ is a function which returns the number of documents in which the words taken as argument appear.

$$C(i, W_i) = \sum_{x=2}^{X} \sum_{y=1}^{x-1} \log \frac{D(w_{i,x}, w_{i,y}) + 1}{D(w_{i,y})} \tag{2.3}$$

In this dissertation, for the models we evaluate, we report the average coherence score of the inferred topics $\frac{1}{K} \sum_{i=1}^{K} C(i, W_i)$. In terms of selecting the number of top words, we vary $X \in \{5, 10, 20\}$ such that the reported coherence scores capture different granularities (i.e.: from a very focused set of words to a more relaxed one). The models which obtain bigger scores are assumed to have more semantically coherent topics.

2.3.2 Document Clustering Evaluation

Document clustering is another form of evaluation for topic models frequently used in the literature [22, 31, 36, 49, 50, 56]. To form the clusters, after the inference procedure, one groups together the documents that have the same topic as the most probable topic in their vector of topic proportions. For example, for a topic model that produces K topics, there are K topic-clusters that can be formed. Say every document has its own topic proportions θ_d. We assign a document d to the topic-

cluster i, when i is the index of $\max_{i \in 1,2,...,K} \theta_{d,i}$.

Given a collection of topic-clusters $\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$, where K is the number of topics produced by a model, and a collection of class-clusters $\mathrm{C} = \{c_1, c_2, ..., c_Z\}$, where Z is the number of ground truth classes and by a class-cluster we understand a group of all the documents which have the same class label, we measure document clustering using standard metrics such as Purity, Normalized Mutual Information and Adjusted Rand Index [28]. These metrics produce scores in the $[0, 1]$ interval, where a higher value means a better performance.

Equation (2.4) lists the formula used to compute *purity*. The idea is to count for each cluster the number of documents in the majority class; then simply divide by the total number of documents M to get a measure of how *pure* the clusters are.

$$\mathrm{Purity}(\Omega, \mathrm{C}) = \frac{1}{M} \sum_{i=1}^{K} \max_{j} |\omega_i \cap c_j| \tag{2.4}$$

The Normalized Mutual Information metric from Equation (2.5) measures the amount of information we obtain about the classes given the clusters and vice-versa, normalized by the entropies of the clusters and classes. The normalization penalizes models which produce a large number of clusters.

$$\begin{aligned} \mathrm{NMI}(\Omega, \mathrm{C}) &= \frac{\mathrm{I}(\Omega, \mathrm{C})}{\frac{\mathrm{H}(\Omega)+\mathrm{H}(\mathrm{C})}{2}} \\ \mathrm{I}(\Omega, \mathrm{C}) &= \sum_{i,j}^{K,Z} \frac{|\omega_i \cap c_j|}{M} \log \frac{M|\omega_i \cap c_j|}{|\omega_i||c_j|} \\ \mathrm{H}(\Omega) &= -\sum_{i=1}^{K} \frac{|\omega_i|}{M} \log \frac{|\omega_i|}{M} \\ \mathrm{H}(\mathrm{C}) &= -\sum_{j=1}^{Z} \frac{|c_j|}{M} \log \frac{|c_j|}{M} \end{aligned} \tag{2.5}$$

The Adjusted Rand Index from Equation (2.6), as its name suggests, is a version of Rand Index whose expected value is 0 (i.e.: corrected for chance). The Rand Index measures clustering in terms of the accuracy of pair-wise decisions - a decision is considered correct if two documents with the same class label are in the same

cluster or if two documents with distinct class labels are in distinct clusters.

$$\mathrm{ARI}(\Omega, \mathrm{C}) = \frac{\sum_{i,j}^{K,Z} \binom{|\omega_i \cap c_j|}{2} - [\sum_{i=1}^{K} \binom{|\omega_i|}{2} \sum_{j=1}^{Z} \binom{|c_j|}{2}]/\binom{M}{2}}{\frac{1}{2}[\sum_{i=1}^{K} \binom{|\omega_i|}{2} + \sum_{j=1}^{Z} \binom{|c_j|}{2}] - [\sum_{i=1}^{K} \binom{|\omega_i|}{2} \sum_{j=1}^{Z} \binom{|c_j|}{2}]/\binom{M}{2}} \tag{2.6}$$

2.3.3 Document Classification Evaluation

Document classification is another extrinsic evaluation task highly utilized in the topic models community. The idea is to use the document-specific topic proportions as the features of the corresponding documents in a classification task. We follow a similar procedure to other researchers [56] and use the Liblinear library [15] from the Weka software [18] with 5-fold cross-validation and the default parameters. We evaluate the document classification performance of the models using Accuracy [28], a metric preferred in many papers from the topic models literature [9, 13, 26, 39, 56].

Let $\{1, 2, ..., M\}$ be a collection of documents where every document d has a true class c_d and a class predicted by the classification algorithm p_d. Accuracy - Equation (2.7) - is defined as the proportion of correct predictions.

$$\text{Accuracy} = \frac{1}{M} \sum_{d=1}^{M} I(c_d = p_d) \tag{2.7}$$

Chapter 3

Topic Models for Single-Context Short Text Data

In short text data topic models are known to suffer from sparsity. The problem is caused by a reduced number of observations available for a reliable inference (i.e.: the words in a document). A popular heuristic utilized to overcome this problem is to perform before training some form of document aggregation by context (e.g.: author, hashtag). The aggregation can alleviate sparsity as the models will be trained on documents with more observations which will also have the potential of being topically related. For example, the publications written by an author will be covering, in most cases, a few if not only one topic (depending on granularity). In this chapter we model explicitly the implicit assumptions of the document aggregation heuristic and apply it to two standard model architectures: a mixture and an admixture. We evaluate the enhanced models in different text environments (i.e.: short and medium) which have different opportunities for aggregation (i.e.: a smaller vs. a bigger number of documents per context). The evaluation targets multiple tasks from topic coherence to document clustering and document classification. Our findings indicate that an admixture model benefits more from aggregation compared to a mixture which rarely improves, and that the state of the art in short text data can be surpassed as long as every context contains a small number of documents.

3.1 Motivation

We know from previous work that in a short text environment, the "one topic per document" assumption of Mixture of Unigrams (MoU) proves to be a good fit to the data, while Latent Dirichlet Allocation (LDA) suffers strongly from sparsity [36, 56]. MoU models global topic proportions whose inference rely on the topics assigned to the documents while LDA assumes document-level topic proportions whose sufficient statistics are the per-word topic assignments. Because short text data items are characterized by a small number of words, the inference of the K-dimensional vector of LDA governing the per-document topic proportions is less reliable (small number of observations).

More than often short text data is accompanied by contextual information such as the date and time of the headline of a news article, the location of a micro-post or the author of the title of a published paper. One popular heuristic employed by researchers to overcome sparsity in short text data is to utilize various message aggregation strategies before training LDA. Hong & Davison [22] find that aggregating tweets based on author gives better performance over standard LDA. In a later paper, Mehrotra *et al.* [31] concludes that aggregating by hashtag brings even more benefits to LDA's performance. These aggregation strategies have also been utilized outside LDA by various researchers who want to avoid sparsity when training their own models on short text data [49, 50, 55].

Motivated by the initial success of document aggregation, in this chapter, we formalize the implicit assumptions the heuristic brings to a topic model and apply it to two standard model architectures: a mixture and an admixture. Concretely, we use context-dependent topic proportions to control the assignment of topics into documents. Documents which share the same context will have their topics drawn according to the same vector of topic proportions. We then extend both LDA (i.e.: the admixture) and MoU (i.e.: the mixture) to accommodate context accompanied text data.

By modelling the implicit assumptions of document aggregation we introduce new building blocks for future and more complex developments. The evaluation aims

to assess whether an admixture model benefits more from aggregation than a mixture model, and how these contextual extensions compare with the state of the art in different text environments (from very short to medium text) and with different amounts of data per context (from a smaller to a bigger opportunity for aggregation). The evaluation targets multiple tasks such as topic coherence, document clustering and document classification.

We want to highlight here TwitterLDA [59] which combines a mixture model with user-specific topic proportions and a background word distribution. TwitterLDA, without the distribution for functional words, becomes an instance of one the models we introduce in this chapter. We argue though, in TwitterLDA, it is unclear whether the performance comes from the fact that a mixture model works good in aggregated short text items or because it models the separation of functional words from content words (the latter point was also made by Vosecky *et al.* [50]). There is also no comparison with the state of the art. In the experiments from this chapter we clearly show that a mixture model trained on aggregated documents does not improve much over its standard version (i.e.: MoU). We can conclude though the reported performance of TwitterLDA is most likely caused by modeling background words in addition to the topic distributions.

We also want to mention that a wide range of topic models has been developed for context accompanied text data. We review a couple of such models in Chapter 2.1.3. These context models vary, but have in common, as the models proposed in this chapter, context-specific topic proportions. The overall difference is that our models have a simpler structure, being built for documents with a single context. Hence, we do not model document-specific preference over contexts as the multi-context topic models do. The purpose behind the models is also distinct: where the multi-context models where defined to capture spatial or temporal topical patterns or to take into account the preferences of the authors for certain topics, our models were defined to formalize the aggregation heuristic utilized in short text data to alleviate sparsity, and to assess whether an admixture benefits more from aggregation compared to a mixture. A further distinction regards the choice of inference. The parameters of the models reviewed in Chapter 2.1.3 are estimated using either EM or Gibbs

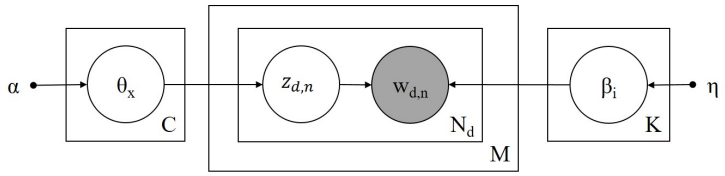


Figure 3.1: Graphical Model of SC-LDA

sampling. In the inference section, we show the models proposed in this chapter are part of a special class of models whose complete conditionals are in the exponential family - this allows both batch and stochastic variational inference to be employed [20], flexibility which can be exploited in both offline and online settings.

3.2 Model Specification

In this section we model explicitly the implicit assumptions of the document aggregation heuristic commonly used to alleviate the sparsity of topic models in short text data. We extend both LDA and MoU to accommodate context accompanied text data. We chose these two models for their set of assumptions (mixture vs. admixture), which make them the standard building blocks for most topic model developments in the literature (see Chapter 2 for a review). We will refer in our discussions to the enhanced models as SC-LDA and SC-MoU.

The models take as input a collection of documents indexed by $d \in \{1, 2, ..., M\}$. Every document d is a collection of words indexed by $n \in \{1, 2, ..., N_d\}$. Every document d is also accompanied by a context c_d. Both models have context-dependent topic proportions which control the assignment of topics into documents. The topic assignments are model-specific: one topic per word for SC-LDA, and one topic per document for SC-MoU.

3.2.1 The SC-LDA Model

The graphical model of SC-LDA is presented in Figure 3.1. The generative process is given below:

- For every topic $i \in \{1, 2, ..., K\}$:
  - Draw a word distribution $\beta_i \sim Dir(\eta)$

- For every context $x \in \{1, 2, ..., C\}$:
  - Draw per-context topic proportions $\theta_x \sim Dir(\alpha)$
- For every document $d \in \{1, 2, ..., M\}$:
  - For every word position $n \in \{1, 2, ..., N_d\}$:
    - Draw a topic $z_{d,n} \sim Cat(\theta_{c_d})$
    - Draw word $w_{d,n} \sim Cat(\beta_{z_{d,n}})$

SC-LDA extends LDA by accommodating contextual information. LDA assumes vectors of document specific topic proportions and one topic per word drawn according to the document level proportions. By defining the context of every document with a unique label the SC-LDA model will degenerate into LDA.

3.2.2 The SC-MoU Model

The graphical model of SC-MoU is presented in Figure 3.2. The generative process is given below:

- For every topic $i \in \{1, 2, ..., K\}$:
  - Draw a word distribution $\beta_i \sim Dir(\eta)$
- For every context $x \in \{1, 2, ..., C\}$:
  - Draw per-context topic proportions $\theta_x \sim Dir(\alpha)$
- For every document $d \in \{1, 2, ..., M\}$:
  - Draw a topic $z_d \sim Cat(\theta_{c_d})$
  - For every word position $n \in \{1, 2, ..., N_d\}$:
    - Draw word $w_{d,n} \sim Cat(\beta_{z_d})$

SC-MoU extends MoU by accommodating contextual information. MoU assumes a vector of global topic proportions and one topic per document drawn according to the global proportions. By defining the context of every document to be the same (call it "global"), the SC-MoU model will degenerate into MoU.

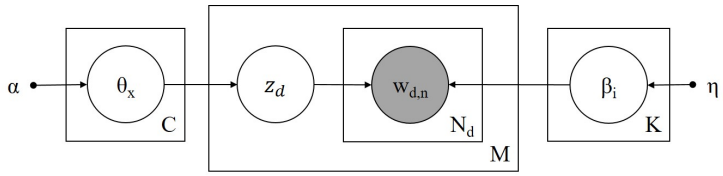


Figure 3.2: Graphical Model of SC-MoU

3.3 Model Inference

To infer the latent parameters of the introduced models, we use the techniques of variational inference for models whose complete conditionals are in the exponential family. Please consult Chapter 2.2.1 for a review. To keep things focused, we give here only an overview of the steps and derivations involved in the inference process - complementing material can be found in Appendix A.

3.3.1 Parameter Inference for SC-LDA

In this section we start by listing the complete conditional of every latent variable of the model given the other latent variables and the observations. Making the observation that each such complete conditional is in the exponential family, we further define the corresponding variational distributions to have the same functional form.

In Equation (3.1) we compute the complete conditional associated with the per-context topic proportions.

$$p(\theta_x|\theta_-, z, \beta, w) = Dir(a), a_i = \alpha_i + \sum_{d,n}^{M,N_d} I(c_d = x)I(z_{d,n} = i) \tag{3.1}$$

Because the complete conditional of the per-context topic proportions is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well $q(\theta_x|\gamma_x) = Dir(\gamma_x)$.

In Equation (3.2) we compute the complete conditional associated with the topics.

$$p(\beta_i|\beta_-, z, \theta, w) = Dir(b), b_j = \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)I(z_{d,n} = i) \tag{3.2}$$

Because the complete conditional of a topic is a Dirichlet, the corresponding

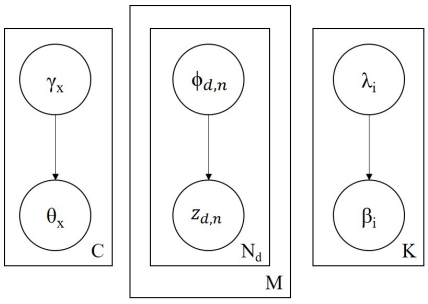


Figure 3.3: The graphical model of the variational distribution used to approximate the posterior of SC-LDA

variational distribution is going to be a Dirichlet as well $q(\beta_i|\lambda_i) = Dir(\lambda_i)$.

In Equation (3.3) we compute the complete conditional associated with the per word topic assignments.

$$p(z_{d,n} = i|z_{-}, \theta, \beta, w) \propto \exp\{\log \theta_{c_d,i} + \log \beta_{i,w_{d,n}}\} \tag{3.3}$$

Because the complete conditional of the per-word topic assignment is a Categorical, the corresponding variational distribution is going to be a Categorical as well $q(z_{d,n}|\phi_{d,n}) = Cat(\phi_{d,n})$.

We have now fully specified the form of the variational distribution used to approximate the posterior of SC-LDA - Figure 3.3 presents its graphical model. Having also specified the complete conditionals, we can derive next the update formulas of the variational parameters. The derivations are made based on the observation that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

In Equation (3.4) we derive the update formula of the variational parameter associated with the topics.

$$\lambda_{i,j} = \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)\phi_{d,n,i} \tag{3.4}$$

In Equation (3.5) we derive the update formula of the variational parameter

associated with the per-context topic proportions.

$$\gamma_{x,i} = \alpha_i + \sum_{d,n}^{M,N_d} I(c_d = x)\phi_{d,n,i} \tag{3.5}$$

In Equation (3.6) we derive the update formula of the variational parameter associated with the per-word topic assignments.

$$\phi_{d,n,i} \propto \exp\{\sum_{x=1}^{C} I(c_d = x)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{j=1}^{V} I(w_{d,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\} \tag{3.6}$$

With the update formulas of the variational parameters at hand, the algorithm is straightforward. The variational parameters are updated iteratively until convergence. This type of algorithm is known in the literature as Coordinate Ascent Mean-Field Variational Inference (CAVI) [10]. Algorithm 1 summarizes one iteration of CAVI.

Algorithm 1 One iteration of Mean Field Variational Inference for SC-LDA.

```
for d = 1 to M do
    for n = 1 to N_d do
        for i = 1 to K do
            Update φ_{d,n,i} using Equation (3.6)
        end for
        Normalize φ_{d,n,*} to sum to 1
    end for
end for
for x = 1 to C do
    for i = 1 to K do
        Update γ_{x,i} using Equation (3.5)
    end for
end for
for i = 1 to K do
    for j = 1 to V do
        Update λ_{i,j} using Equation (3.4)
    end for
end for
```

3.3.2 Parameter Inference for SC-MoU

The inference steps are similar to the ones taken in Section 3.3.1 for SC-LDA. Since the vector of per-context topic proportions is a particularity for both SC-LDA and

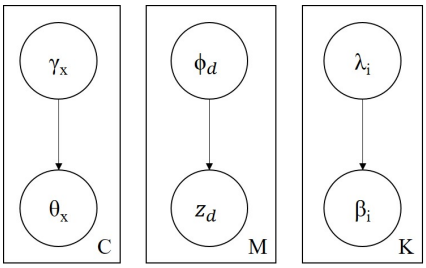


Figure 3.4: The graphical model of the variational distribution used to approximate the posterior of SC-MoU

SC-MoU, the associated complete conditional (also its corresponding variational distribution) is the same as in Equation (3.1).

In Equation (3.7) we compute the complete conditional associated with the topics.

$$p(\beta_i|\beta_-, z, \theta, w) = Dir(b), b_j = \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)I(z_d = i) \tag{3.7}$$

Because the complete conditional of a topic is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well $q(\beta_i|\lambda_i) = Dir(\lambda_i)$.

In Equation (3.8) we compute the complete conditional associated with the per-document topic assignments.

$$p(z_d = i|z_-, \theta, \beta, w) \propto \exp\{\log \theta_{c_d,i} + \sum_{n=1}^{N_d} \log \beta_{i,w_{d,n}}\} \tag{3.8}$$

Because the complete conditional of the per-document topic assignment is a Categorical, the corresponding variational distribution is going to be a Categorical as well $q(z_d|\phi_d) = Cat(\phi_d)$.

Figure 3.4 presents the graphical model of the variational distribution used to approximate the posterior of SC-MoU.

Having computed the complete conditionals and having defined the form of the corresponding variational distributions, we can now derive the update formulas of the variational parameters.

In Equation (3.9) we derive the update formula of the variational parameter

associated with the topics.

$$\lambda_{i,j} = \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)\phi_{d,i} \tag{3.9}$$

In Equation (3.10) we derive the update formula of the variational parameter associated with the context topic proportions.

$$\gamma_{x,i} = \alpha_i + \sum_{d=1}^{M} I(c_d = x)\phi_{d,i} \tag{3.10}$$

In Equation (3.11) we derive the update formula of the variational parameter associated with the per-document topic assignments.

$$\phi_{d,i} \propto \exp\{\sum_{x=1}^{C} I(c_d = x)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{n,j}^{N_d,V} I(w_{d,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\} \tag{3.11}$$

With the update formulas of the variational parameters at hand, the algorithm used for inference is similar to the one already introduced in Algorithm 1.

3.3.3 Document-level Topic Proportions

The introduced topic models do not model directly document-level topic proportions. In fact, we specifically avoid doing that in order to alleviate sparsity. Having a topical representation of a document is nevertheless useful for both summarizing the document and as a feature in many tasks such as clustering and classification.

We compute the document-level topic proportions of the models using the available sufficient statistics: the per-word topic assignments in case of SC-LDA - Equation (3.12) - and the per-document topic assignments for SC-MoU - Equation (3.13).

$$p(topic = i|d) \propto \alpha_i + \sum_{n=1}^{N_d} \phi_{d,n,i} \tag{3.12}$$

$$p(topic = i|d) \propto \alpha_i + \phi_{d,i} \tag{3.13}$$

The topical representation of the documents is somewhat ill-defined in work

which uses the document-aggregation heuristic applied to LDA - for simplicity, researchers use the topic proportions associated with the macro-document to represent the documents part of it [22, 31]. Translated into the formalism included in this chapter, it would mean documents part of the same context have the same topical representation. One can clearly understand why this is not the case. The topic proportions of a context give an overall information for all the documents part of it - so individual documents should have their own topical representation (which can happen to be different).

3.4 Evaluation

We evaluate the model on four labeled datasets of scientific publications. The datasets cover multiple text environments (short and medium) and have different opportunities for aggregation (a smaller vs. a larger number of documents per contexts). The evaluation targets multiple tasks such as topic coherence, document clustering and document classification. Please consult Chapter 2.3 for details about the metrics utilized to assess these tasks.

The following models are used in the evaluation for comparison:

- **Latent Dirichlet Allocation (LDA)** This baseline corresponds to the SC-LDA model where every document has its own context. We use the prior values recommended in previous work [36, 56] ($\alpha = 0.1$; $\eta = 0.01$). For a review of this model please consult Chapter 2.1.1.2.

- **SC-LDA-FA** This is the SC-LDA model with first author as context. Common values for sparse priors are used ($\alpha = 0.1$; $\eta = 0.01$).

- **Mixture of Unigrams (MoU)** This baseline corresponds to the SC-MoU model where all documents share the same context. We use the prior values recommended in previous work [56] ($\alpha = 50/K$; $\eta = 0.01$). For a review of this model please consult Chapter 2.1.1.1.

- **SC-MoU-FA** This is the SC-MoU model with first author as context. Common values for sparse priors are used ($\alpha = 0.1$; $\eta = 0.01$).

- **Biterm Topic Model (BTM)** This model proposed by Yan *et al.* [56] has been selected to represent the state of the art in short text data. The model has a preprocessing step in which all the biterms (i.e.: word pairs) of every document are generated. Then the biterms become the input of a Mixture of Unigrams model (i.e.: global topic proportions and one topic per biterm). For a fair comparison, we reimplemented the model with Variational Inference (original implementation is done using Gibbs Sampling). We use the same priors as in the original paper ($\alpha = 50/K$; $\eta = 0.01$).

The models are initialized according to standard practices from the literature. Blei & Lafferty [8] find that a good way to initialize the topics is to use a random sample of N documents from the corpus and compute a smoothed word distribution over the vocabulary space from the word counts of the random sample. We choose N to be 10.

We perform the evaluation with 3 levels of K (i.e.: the number of topics): $K = Z$, $K = 2Z$ and $K = 3Z$, where Z is the number of ground truth classes of a dataset. For each setting of a model we do 10 runs and report the result that has the maximum ELBO - the bigger the ELBO the closer the variational approximation is to the true posterior.

3.4.1 Dataset Selection

We use in the evaluation four datasets of scientific publications downloaded from arXiv (www.arxiv.org), a well known digital library. Two of the datasets contain titles of publications and represent the short text environment; while the other two contain the abstracts from almost (due to preprocessing) the same set of publications - the medium text environment. For both the short and medium text environments we have one dataset with a small number of documents per context (average of 8) and another with a larger number (average of 28). We refer to the created datasets as “Short Text Small Contexts”, “Short Text Larger Contexts”, “Medium Text Small Contexts” and “Medium Text Larger Contexts”. The datasets were created to facilitate an assessment of how different opportunities for aggregation affect the

Dataset	Classes	Documents	Unique words	All words	Average words per document
Short Text Small Contexts	10	20000	6696	139702	6.986
Short Text Larger Contexts	10	14587	2833	92518	6.342
Medium Text Small Contexts	10	20005	11991	1463973	73.180
Medium Text Larger Contexts	10	16501	5276	1074094	65.093

Table 3.1: Statistics for the datasets used in the evaluation

performance of the models in different text environments.

To avoid any bias we selected the publications from a single big subject (i.e.: physics). The type of bias we wanted to avoid is to have mixed subjects (e.g.: physics + biology + literature) where it is almost guaranteed that authors would not cross-publish. This type of bias arises especially in papers that target tweets, where ground truth labels are usually not available. For example, in the work of Mehrotra *et al.* [31], the input collections are constructed by making queries to a large sample of tweets, and labeling the documents with the query terms that retrieved them (e.g.: "music", "food", "sport"). Another relevant example is the work of Hong & Davison [22], where the researchers use the categories assigned to the users as the ground truth labels for their messages. They explicitly specify there is no overlap between the categories. We gave these two specific examples because they are the ones in which various document aggregation techniques are evaluated for LDA and were part of the motivation behind this chapter. When the authors do not cross-publish, the contexts become clear ground truth discriminators. Datasets in which contexts contain documents from more than one ground truth class should increase the difficulty of the evaluation.

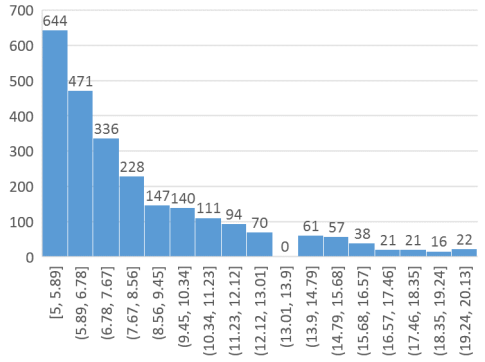
Another reason for selecting this type of data (i.e.: titles and abstracts of scientific publications) is the reduced noise; so any topic model which will be trained on it will be able to bring out the performance of its generative process without being overloaded by high frequency, non-topical words (as it happens for example in a dataset of tweets).

Dataset	First Authors	Average documents per first author
Short Text Small Contexts	2477	8.074
Short Text Larger Contexts	523	27.891
Medium Text Small Contexts	2462	8.126
Medium Text Larger Contexts	585	28.207

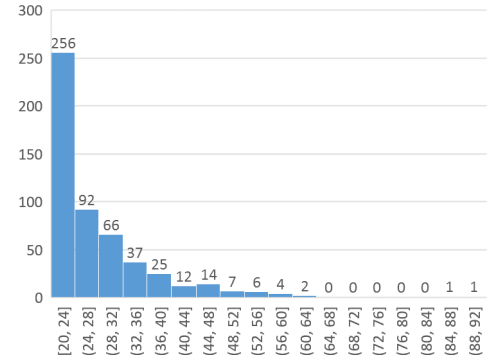
Table 3.2: First Author statistics for the datasets

Table 3.1 summarizes useful dataset statistics after preprocessing (basic stop and rare word removal). The datasets constructed to represent the short text environments have an average of 7 words per document, while the ones for medium text have 65 and 73, respectively. Another useful statistic is that there are 10 ground truth classes. The labels correspond to different areas from physics (e.g.: "Condensed Matter", "Nuclear Theory"). In Table 3.2 one can find the statistics associated with the "First Author" context for every dataset. In the datasets with small contexts (characterized by an average of 8 documents) there is a much larger number of first authors (2400+ vs. 500+) compared to the larger contexts datasets (characterized by an average of 28 documents). In Figure 3.6 we show how many authors published in one or more ground truth categories. We can see that the larger context datasets enjoy a larger spread. This can be explained by the fact that these datasets have an increased number of documents per context (hence a bigger opportunity to cross-publish). For completeness, Figure 3.5 shows the distribution of the number of documents per first authors for every dataset.

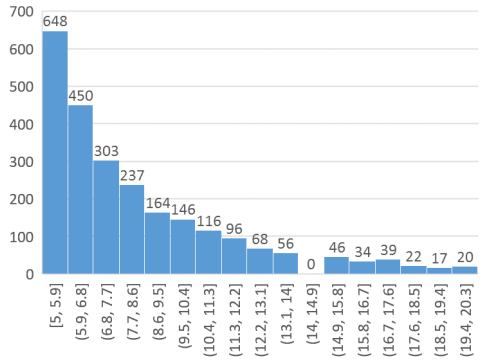
We will now give some details which facilitate the replicability of the experimental setup. First, we list all the ground truth categories in Table 3.3. The arXiv contains 13 categories from physics, but 3 of them had a small number of documents and ended up being discarded. As a general rule, we considered only the documents which belong to a single category. In order to build the datasets with small contexts, we chose only those documents which belong to authors that have between 5 and 20 publications (see the distributions in the first column of Figure 3.5). Because the size of the resulted datasets was too large for batch variational inference, we



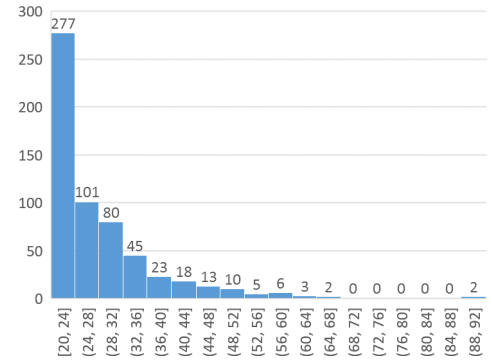
(a) "Short Text Small Contexts" dataset



(b) "Short Text Larger Contexts" dataset



(c) "Medium Text Small Contexts" dataset



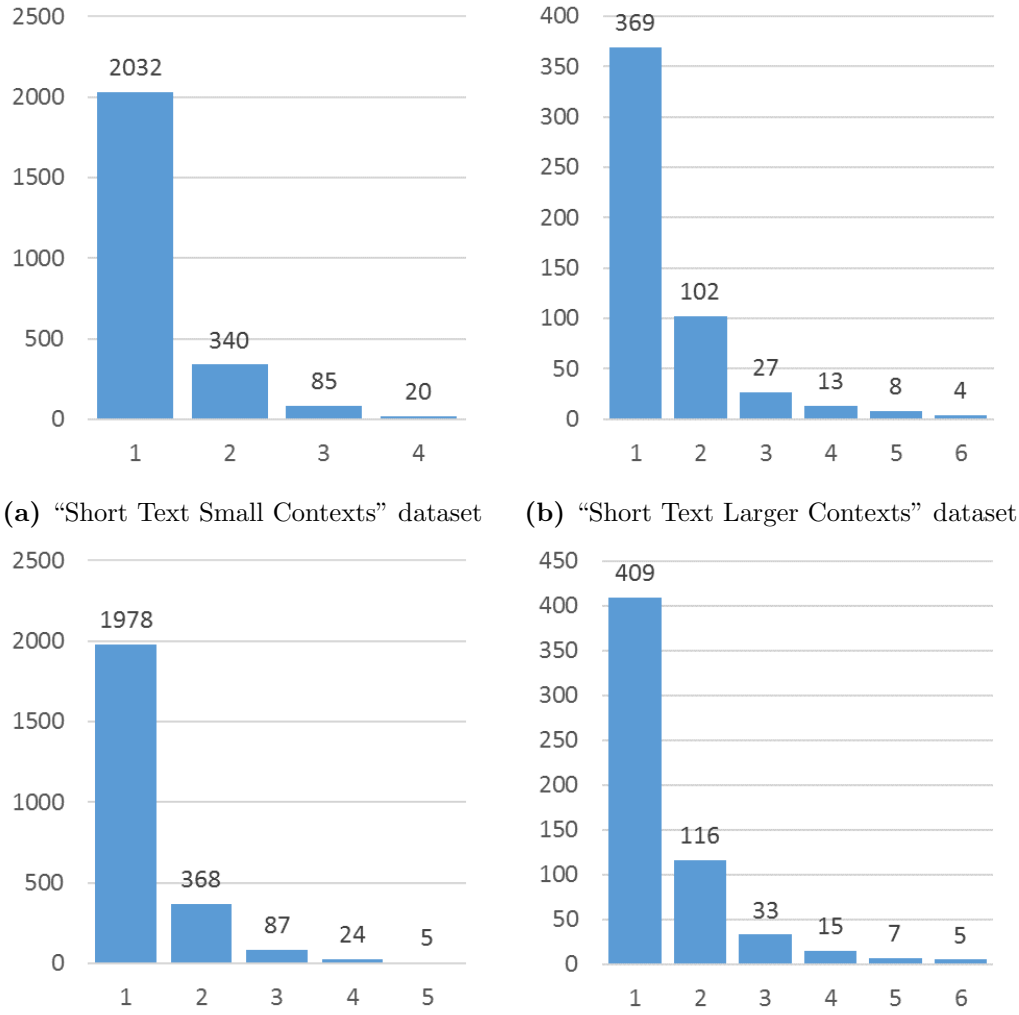
(d) "Medium Text Larger Contexts" dataset

Figure 3.5: The distribution of the number of documents per first authors. The vertical axis shows the number of authors; the horizontal axis lists intervals of numbers of documents.

Unique arXiv identifier	Description
physics:gr-qc	General Relativity and Quantum Cosmology
physics:astro-ph	Astrophysics
physics:quant-ph	Quantum Physics
physics:hep-lat	High Energy Physics - Lattice
physics:cond-mat	Condensed Matter
physics:nucl-th	Nuclear Theory
physics:nlin	Nonlinear Sciences
physics:hep-th	High Energy Physics - Theory
physics:hep-ph	High Energy Physics - Phenomenology
physics:physics	Physics (other)

Table 3.3: The arXiv categories utilized to construct the datasets

sampled randomly approximately 20,000 documents. To build the datasets with larger contexts, we considered only the documents which belong to authors that have more than 20 publications (see the distributions in the second column of Figure 3.5). Since the resulting datasets had reasonable sizes, we kept them as they were. In terms of preprocessing, we removed stop words, words with a length smaller than 3, words with a global frequency smaller than 5 for the title datasets and smaller than 20 for the abstract datasets. We also discarded the documents which had less than 3 words inside and those which belong to authors with ambiguous names (e.g.: only letters given; we discarded authors that had the surname or the forenames smaller than 3 letters).



(a) "Short Text Small Contexts" dataset **(b)** "Short Text Larger Contexts" dataset

(c) "Medium Text Small Contexts" dataset **(d)** "Medium Text Larger Contexts" dataset

Figure 3.6: The number of first authors (vertical axis) that published documents in one or more ground truth classes (horizontal axis).

Dataset	Model	Top 5 words	Top 10 words	Top 20 words
Short Text Small Contexts	LDA	-22.866	-140.049	-684.379
	SC-LDA-FA	-19.155	-128.452	-647.265
	MoU	-19.719	-126.775	-640.453
	SC-MoU-FA	-19.935	-124.843	-636.639
	BTM	**-17.304**	**-119.244**	**-623.205**
Short Text Larger Contexts	LDA	-21.803	-131.204	-680.870
	SC-LDA-FA	**-17.628**	**-118.817**	-618.784
	MoU	-19.035	-125.617	-627.604
	SC-MoU-FA	-18.214	-120.702	-624.448
	BTM	-18.206	-119.148	**-606.940**
Medium Text Small Contexts	LDA	**-10.259**	-69.486	-361.202
	SC-LDA-FA	-10.457	**-67.579**	**-355.795**
	MoU	-13.604	-86.382	-408.539
	SC-MoU-FA	-15.099	-93.633	-443.087
Medium Text Larger Contexts	LDA	-10.403	-68.001	**-352.820**
	SC-LDA-FA	**-10.058**	**-65.566**	-353.535
	MoU	-11.573	-72.316	-414.466
	SC-MoU-FA	-11.573	-72.916	-414.039

Table 3.4: Topic Coherence results with K set to the number of ground truth classes (K=10).

3.4.2 Topic Coherence Evaluation

In this section we present and discuss the results for topic coherence, a measure of topic quality which aims to capture the human interpretability of topics in an automatic fashion (i.e.: no human annotators). Please consult Chapter 2.3.1 for a review of the task and details about the utilized metric.

In Table 3.4 we list the topic coherence results when K is set to the number of ground truth classes, whereas in Table 3.5 we show the trends when K varies. On the short text datasets, LDA is the worst performing model across all levels of K and number of top words, confirming that sparsity drastically affects this model (see the first two rows of Table 3.5). On short text data with small contexts (first row of Table 3.5), SC-LDA-FA is second to last, managing to outperform overall only its non-context baseline (i.e.: LDA) - note though that there is a clear improvement in performance brought by the context, but still, the model does not shine compared to the top performers. When the contexts contain more opportunity for aggregation and K is set to the number of ground truth classes (see second row of Table 3.4), SC-LDA-FA is overall the best model. However, as K is increasing, sparsity starts

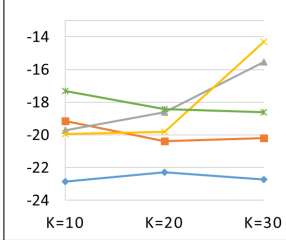
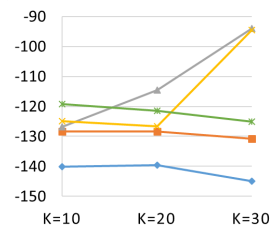
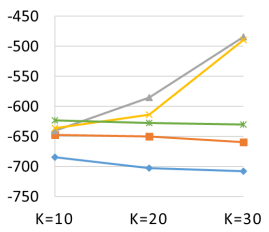
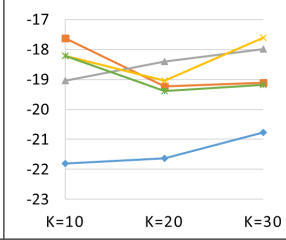
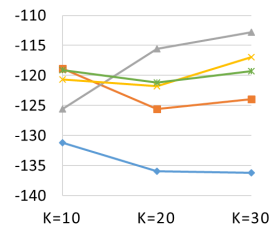
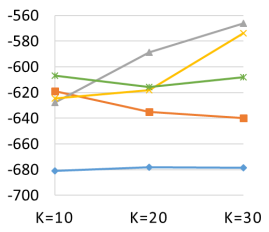
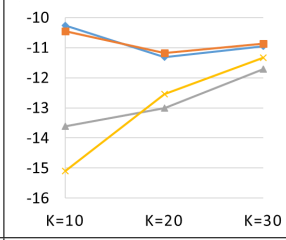
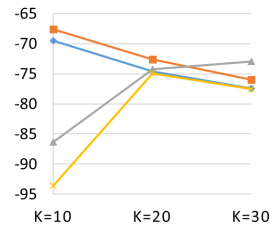
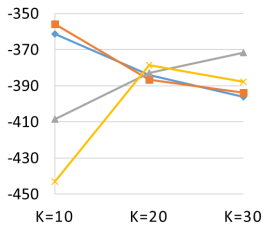
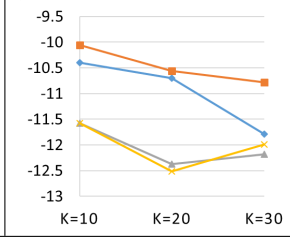
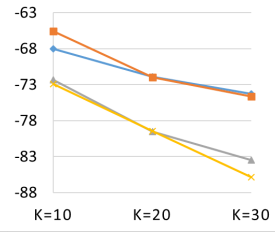
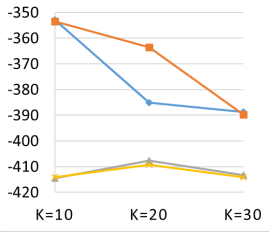
Dataset	Top 5 words	Top 10 words	Top 20 words
Short Text Small Contexts			
Short Text Larger Contexts			
Medium Text Small Contexts			
Medium Text Larger Contexts			

Table 3.5: Topic Coherence results when K varies

to reappear, and the performance of SC-LDA-FA drops again to the penultimate place (second row of Table 3.5). In general, on the short text datasets (first 2 rows of Table 3.5), MoU is the better model; its context-extended counterpart, SC-MoU-FA, fails to register stable improvements. BTM, our choice for the state of the art, outperforms the other models only when K is set to the number of ground truth classes (first 2 rows of Table 3.4). As K starts to increase, BTM's performance drops below those of "one topic per document" models (best seen in the first row of Table 3.5).

On medium-sized text, the worst performing models are MoU and SC-MoU-FA (last 2 rows of Tables 3.4 and 3.5) - this is in accordance with the expectation

Dataset	Model	Purity	NMI	ARI
Short Text Small Contexts	LDA	0.550	0.190	0.113
	SC-LDA-FA	**0.784**	**0.499**	**0.361**
	MoU	0.658	0.339	0.325
	SC-MoU-FA	0.664	0.354	0.342
	BTM	0.769	0.455	**0.361**
Short Text Larger Contexts	LDA	0.578	0.233	0.195
	SC-LDA-FA	**0.820**	**0.584**	0.462
	MoU	0.752	0.489	0.546
	SC-MoU-FA	0.789	0.543	**0.579**
	BTM	0.805	0.533	0.448
Medium Text Small Contexts	LDA	0.770	0.502	0.348
	SC-LDA-FA	**0.824**	**0.587**	**0.400**
	MoU	0.588	0.370	0.365
	SC-MoU-FA	0.589	0.374	0.370
Medium Text Larger Contexts	LDA	0.819	0.551	0.408
	SC-LDA-FA	**0.886**	**0.686**	**0.592**
	MoU	0.730	0.590	0.566
	SC-MoU-FA	0.734	0.605	0.577

Table 3.6: Document Clustering results with K set to the number of ground truth classes (K=10).

that the "one topic per document" assumption of these models is unsuitable for any piece of text that is not short. LDA becomes competitive now since there are more observations per documents available, while SC-LDA-FA is overall the best performing model.

3.4.3 Document Clustering Evaluation

In this section we present and discuss the results for document clustering. We asses this task with three common metrics: Purity, Normalized Mutual Information and Adjusted Rand Index. Please consult Chapter 2.3.2 for a review of the task and details about the utilized metrics.

When the performance of the document clustering task is measured in terms of Purity and NMI, SC-LDA-FA is clearly the best performing model across all evaluated text environments and levels of K (Tables 3.6 and 3.7). Sticking with the same metrics, BTM, our choice for the state of the art, comes second place (first 2 rows of Tables 3.6 and 3.7). LDA is the worst performer across all metrics, reconfirming the drastic impact sparsity has on this model - result more pronounced

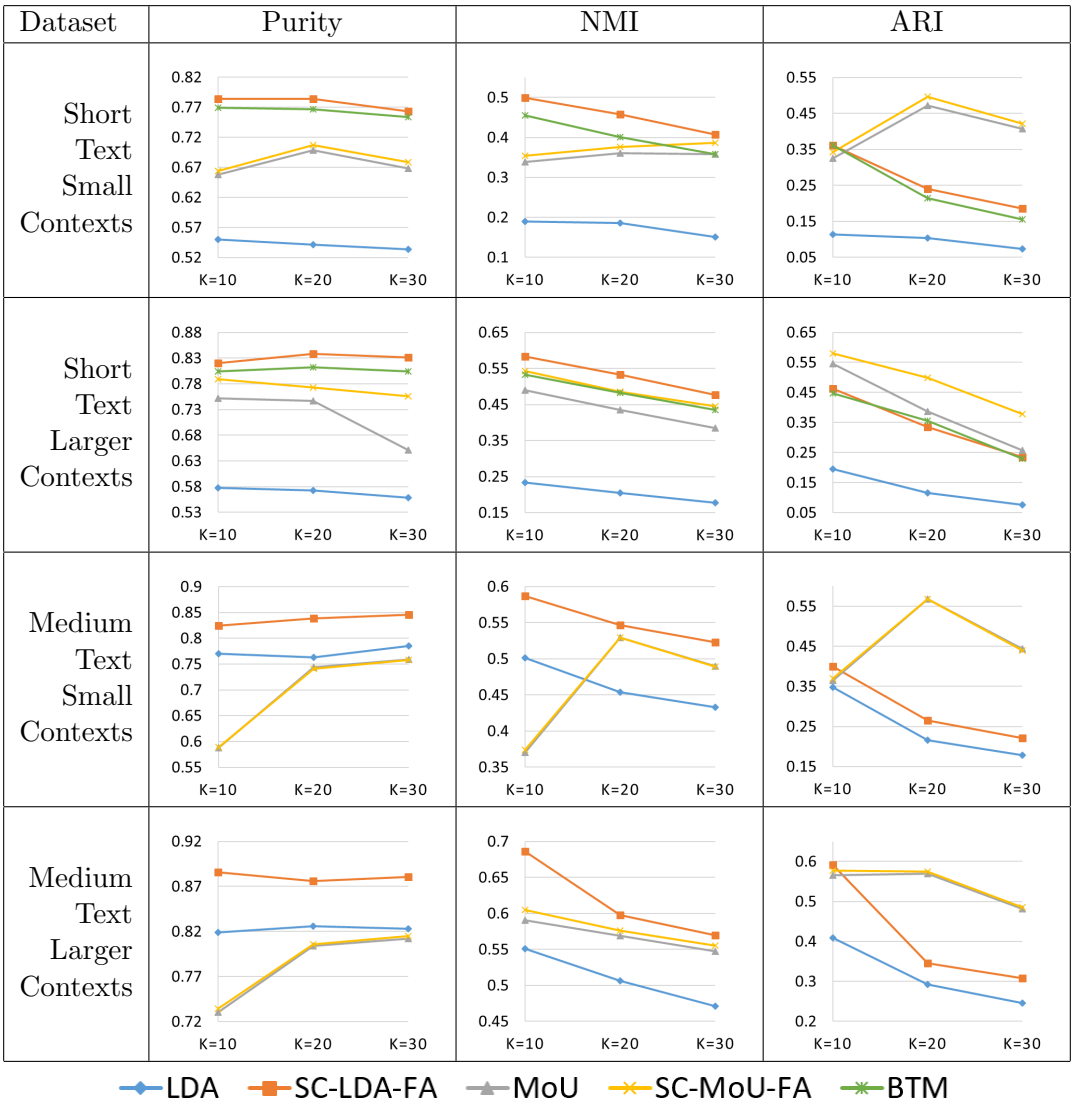


Table 3.7: Document Clustering results when K varies

on the short text datasets but present also in the medium text datasets. The "one topic per document" models, MoU and SC-MoU-FA, are performing in most cases on par suggesting that context is not helpful in this evaluation setting. Another observation about the mixture models is that they perform overall better in terms of ARI than all the other models - a possible explanation might be that their hard constraint helps them reduce the false positive and false negative clustering decisions penalized by the metric.

3.4.4 Document Classification Evaluation

In this section we present and discuss the results for document classification. We asses this task in terms of Accuracy. Please consult Chapter 2.3.3 for a review of the task and details about the utilized metric.

For the document classification task the results indicate clear and consistent rankings across all levels of K. On the short text datasets (first row of Table 3.8), SC-LDA-FA is the best performing model. The second best model is BTM. The ranking is completed by SC-MoU-FA, MoU and LDA. LDA is by far the worst performer, making the gains obtained by SC-LDA-FA reach very large values (e.g. from Table 3.8: 0.55 vs 0.82; 0.58 vs. 0.84). This is another reconfirmation of the drastic effect sparsity has on LDA. We make the observation that for a classification task, a model like LDA, which assumes one topic per word, can produce more features for document representation (a maximum of K) than MoU or SC-MoU-FA (which produce only one feature with high mass). Nevertheless, LDA, because of sparsity, ranks below these "one topic per document" models. We further point out that the context extension of MoU, SC-MoU-FA obtained a good performance boost over the standard mixture model.

On the medium text datasets, SC-LDA-FA is again the best model. Since the length of documents is far bigger now than in the previous datasets (see Table 3.1 for the statistics), LDA becomes the second best model, surpassing both MoU and SC-MoU-FA which seem to be performing on par.

3.5 Discussion

In this chapter we explicitly modelled the implicit assumptions of document aggregation, a popular heuristic employed to alleviate the sparsity suffered by topic models in short text environments, and applied it to two standard model architectures: a mixture and an admixture. We evaluated the enhanced models on both very short (i.e.: titles of publications) and medium (i.e.: abstracts) text items, with different opportunities for aggregation (a smaller vs. a larger number of documents per context). Since the target was short text data, we included for comparison a state of

Dataset	Model	Accuracy
Short Text Small Contexts	LDA	0.555
	SC-LDA-FA	**0.822**
	MoU	0.655
	SC-MoU-FA	0.664
	BTM	0.790
Short Text Larger Contexts	LDA	0.583
	SC-LDA-FA	**0.840**
	MoU	0.732
	SC-MoU-FA	0.790
	BTM	0.818
Medium Text Small Contexts	LDA	0.812
	SC-LDA-FA	**0.863**
	MoU	0.638
	SC-MoU-FA	0.639
Medium Text Larger Contexts	LDA	0.845
	SC-LDA-FA	**0.915**
	MoU	0.721
	SC-MoU-FA	0.734

Table 3.8: Document Classification results with K set to the number of ground truth classes (K=10).

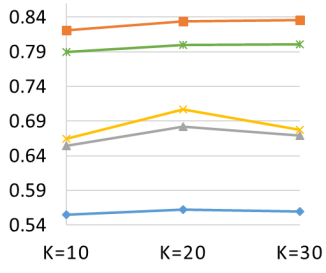
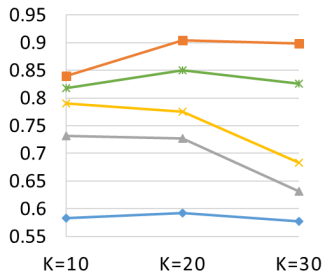
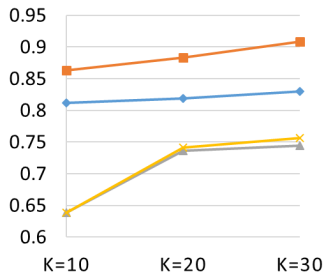
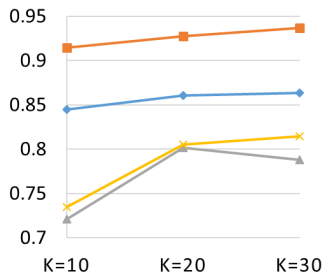
Dataset	Accuracy	Dataset	Accuracy
Short Text Small Contexts		Short Text Larger Contexts	
Medium Text Small Contexts		Medium Text Larger Contexts	

Table 3.9: Document Classification results when K varies

the art model from this area of topic modelling [56]. The evaluation assessed topic coherence, document clustering and document classification. We list below the main findings:

- **Clustering** The context extension of LDA is the best performing model across the evaluated datasets (short and medium text; smaller and larger contexts), surpassing the state of the art in short text data. The context extension of MoU brings little to no benefits over the standard mixture model.
- **Classification** We find that the context enhanced version of LDA outperformed the state of art on short text (in both cases of smaller and larger contexts). The context extension of MoU brings good improvements over its standard version only on short text data with larger contexts; even so, the model is far inferior to the context extension of LDA, also ranking below the state of the art. The context extension of LDA is, at the same time, the best performer on the medium text datasets.
- **Topic Coherence** We find there is no clear generic pattern that favours one model to the other here. Please consult Section 3.4.2 for more fine grained patterns.

Based on the assessments made in this chapter, we can conclude that the context extension of LDA is overall the best performing model, capable of surpassing the state of the art in short text data when there is at least a small amount of aggregation available for each context. The model assumptions of LDA are also the ones most probable to benefit from aggregation. We find that the context extension of MoU rarely improves over the standard mixture model. "Statistically" this makes sense. In both cases (the context extension of LDA and MoU) one has to infer context specific topic proportions. In case of LDA, the "one topic per word" assumption produces many more sufficient statistics compared with MoU and its "one topic per document" assumption. This means, for an admixture, we have more confidence in the inference of the K-dimensional vector of context topic proportions.

We would also like to discuss some limitations behind the work from this chapter. First of all, we did not intend to identify which context is most suitable for

aggregation (performance-wise). Previous studies, which were also the motivation behind this chapter, already showed, for example, that author or hashtag are useful context choices [22, 31]. Instead, in this chapter, we gave a formal treatment to the document aggregation heuristic applied to topic models and built an experimental set up that allowed us to determine which class of models - a mixture vs. an admixture - benefits more from aggregation. A secondary objective was to assess whether document aggregation can lead to state of the art performance. Nevertheless, we note that we also experimented with other choices of context which provided results below the ones reported in this chapter. We noticed that contextual information is not useful when it is shared by documents that are not topically related. For example, utilizing "month" as context, would not be suitable, as papers from all the ground truth classes can be published in a certain month. This observation applies to our datasets, but can be quite inadequate for others where the "month" context can happen to be a good topical discriminator. This warrants further future work to show how this will generalize beyond the chosen datasets and context.

Chapter 4

A Co-occurrence-based Topic Model for Short Text Data

In normal text data, the availability of repeating word co-occurrences across the documents is known as a core contributor to the discovery of latent topics. However, the possibility of such patterns to arise in a short text dataset is reduced. With this observation in mind, we propose a new model for short text data which assumes every document is a bag of word co-occurrences, where each co-occurrence belongs to a latent topic. The documents are enhanced a priori with related co-occurrences from the other documents, such that the collection will have a greater chance than before to exhibit word patterns. We evaluate the model on two labeled datasets of tweets and one of titles of scientific publications. The evaluation targets multiple tasks such as topic coherence, document clustering and document classification. The proposed model performs well managing to surpass the state of the art and popular topic model baselines.

4.1 Motivation

We have previously addressed in detail (i.e.: Chapter 3) the reasons why models like Latent Dirichlet Allocation (LDA) fail on short text data, and how simpler models like Mixture of Unigrams (MoU) manage to obtain a better performance. The approach we took to alleviate sparsity was to exploit the available context which

accompanies certain types of short text data items (e.g.: the author in a dataset of titles of scientific publications). However, contextual information is not always available or it does not help (i.e.: it is shared by documents which have little or no topical relationship). For these types of situations a general purpose model for short text data is desirable.

Admixture models like LDA work well on normal text collections because word co-occurrence patterns arise across the documents. Because of the small number of words per document that characterizes short text collections, the opportunity for such patterns to arise consistently in this environment is reduced. With this observation in mind, we propose a new model for short text data which assumes that every document is a bag of word co-occurrences, where each co-occurrence belongs to a latent topic. The documents are enhanced a priori with related co-occurrences from the other documents, such that the collection will have increased chances of exhibiting word patterns.

We evaluate the proposed model on two labeled datasets of tweets and one of titles of scientific publications. The latter is a dataset we previously used in the evaluation from Chapter 3, where the context extension of LDA (i.e.: SC-LDA-FA) was, overall, the best performer. We introduce this dataset in the evaluation because, in addition to comparing the model with popular topic baselines and the state of the art, we want to assess its performance relative to a model like SC-LDA-FA which leverages contextual information. This allows us to assess whether utilizing context (when available) still leads to better results. The evaluation targets multiple tasks such as topic coherence, document clustering and document classification.

In the inference process, we show the model is part of a special class of models whose complete conditionals are in the exponential family - this allows both batch and stochastic variational inference to be employed [20], flexibility which can be exploited in both offline and online settings.

We want to highlight here the Biterm Topic Model (BTM) of Yan *et al.* [56]. The model makes use of co-occurrences in the form of biterms (i.e.: a pair of words). It builds a collection of all the biterms which can be generated from the documents taken as input. BTM assumes global topic proportions to alleviate sparsity and

one topic per biterm which leads to a richer model compared to MoU (as documents have more than one biterm; hence the possibility to exhibit more topics). The model proposed in this chapter will also use "biterms" as a choice for word co-occurrences due to their simplicity, but the work is quite different: our model assumes document specific topic proportions and reduces sparsity by enhancing each document with relevant "biterms" from the collection. We note that BTM does not have a parameter which directly captures the topical representation of the documents. The authors do provide though a way of indirectly calculating this vector of probabilities with the parameters of the model. In the evaluation we find our model to outperform BTM.

4.2 Model Specification

In this section we describe a new topic model for short text data which is based on word co-occurrences. We will refer in our discussions to the proposed model as CTM (Co-occurrence Topic Model).

The model takes as input a collection of documents indexed by $d \in \{1, 2, ..., M\}$. Every document d is a collection of word co-occurrences indexed by $p \in \{1, 2, ..., N_d\}$. Every co-occurrence p is a collection of words indexed by $n \in \{1, 2, ..., N_{d,p}\}$.

The graphical model of CTM is presented in Figure 4.1. The generative process is given below:

- For every topic $i \in \{1, 2, ..., K\}$:
  - Draw a word distribution $\beta_i \sim Dir(\eta)$
- For every document $d \in \{1, 2, ..., M\}$:
  - Draw document-level topic proportions $\theta_d \sim Dir(\alpha)$
  - For every word co-occurrence $p \in \{1, 2, ..., N_d\}$:
    - Draw a topic $z_{d,p} \sim Cat(\theta_d)$
    - For every word position $n \in \{1, 2, ..., N_{d,p}\}$:
      - Draw word $w_{d,p,n} \sim Cat(\beta_{z_{d,p}})$

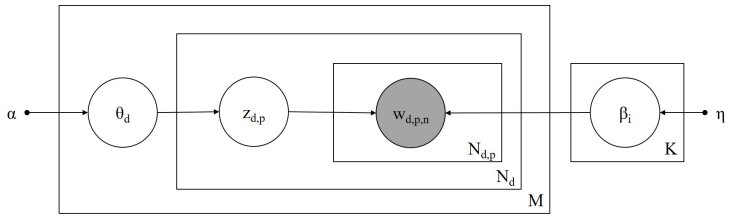


Figure 4.1: Graphical Model of CTM

The model requires the collections associated with the word co-occurrences from every document to be defined before the learning phase. The generative process expressed above gives a formal description of the model in a generic scenario. In our experiments we focus on one special case of word co-occurrences, those formed of two words only. For every document we compute all the word pairs. We then build a global pair co-occurrence matrix (where each entry tells the frequency of co-occurrence between two word pairs). Finally, for each pair that belongs to a document we extract from the global matrix the top T pairs. Now every document will have the original word pairs plus the ones we just selected. Because of the large overlap of pair co-occurrences and noise, we choose to simply represent the document as the set of the pairs that result from the described selection process. Future work can look into better ways of selecting word co-occurrences that are related to a document (e.g.: taking into consideration, besides frequency, the coverage across the documents; or utilizing n-gram co-occurrences). The idea was to simply enhance the documents with related co-occurring words such that patterns arise across the collection - this is similar to what happens in a normal text environment in the case of a admixture model like LDA.

We note that CTM can be viewed as a general extension of LDA. When every word from a document is placed into a single-element co-occurrence collection, the model degenerates into LDA.

4.3 Model Inference

To infer the latent parameters of the introduced model, we use the techniques of variational inference for models whose complete conditionals are in the exponential family. Please consult Chapter 2.2.1 for a review. To keep things focused, we give

here only an overview of the steps and derivations involved in the inference process - complementing material can be found in Appendix B.

We start by listing the complete conditional of every latent variable of the model given the other latent variables and the observations. Making the observation that each such complete conditional is in the exponential family, we further define the corresponding variational distributions to have the same functional form.

In Equation (4.1) we compute the complete conditional associated with the per-document topic proportions.

$$p(\theta_d|\theta_-, z, \beta, w) = Dir(a), a_i = \alpha_i + \sum_{p=1}^{N_d} I(z_{d,p} = i) \tag{4.1}$$

Because the complete conditional of the per-document topic proportions is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well $q(\theta_d|\gamma_d) = Dir(\gamma_d)$.

In Equation (4.2) we compute the complete conditional associated with the topics.

$$p(\beta_i|\beta_-, z, \theta, w) = Dir(b), b_j = \eta_j + \sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n} = j)I(z_{d,p} = i) \tag{4.2}$$

Because the complete conditional of the topics is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well $q(\beta_i|\lambda_i) = Dir(\lambda_i)$.

In Equation (4.3) we compute the complete conditional associated with the per word co-occurrence topic assignments.

$$p(z_{d,p} = i|z_-, \theta, \beta, w) \propto \exp\{\log \theta_{d,i} + \sum_{n=1}^{N_{d,p}} \log \beta_{i,w_{d,p,n}}\} \tag{4.3}$$

Because the complete conditional of the per word co-occurrence topic assignment is a Categorical, the corresponding variational distribution is going to be a Categorical as well $q(z_{d,p}|\phi_{d,p}) = Cat(\phi_{d,p})$.

We have now fully specified the form of the variational distribution used to approximate the posterior of CTM - Figure 4.2 presents its graphical model. Having

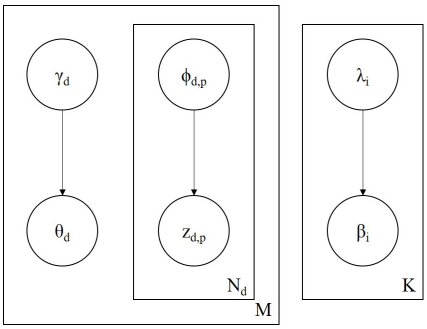


Figure 4.2: The graphical model of the variational distribution used to approximate the posterior of CTM

also specified the complete conditionals, we can derive next the update formulas of the variational parameters. The derivations are made based on the observation that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

In Equation (4.4) we derive the update formula of the variational parameter associated with the topics.

$$\lambda_{i,j} = \eta_j + \sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n} = j)\phi_{d,p,i} \tag{4.4}$$

In Equation (4.5) we derive the update formula of the variational parameter associated with the per-document topic proportions.

$$\gamma_{d,i} = \alpha_i + \sum_{p=1}^{N_d} \phi_{d,p,i} \tag{4.5}$$

In Equation (4.6) we derive the update formula of the variational parameter associated with the per word co-occurrence topic assignments.

$$\phi_{d,p,i} \propto \exp\{\Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0}) + \sum_{n,j}^{N_{d,p},V} I(w_{d,p,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\} \tag{4.6}$$

With the update formulas of the variational parameters at hand, the algorithm is straightforward. The variational parameters are updated iteratively until convergence. This type of algorithm is known in the literature as Variational EM [9]. The pseudo-code can be found in Algorithm 2.

Algorithm 2 Variational EM for CTM

```
Initialize λ
while no global convergence do
    for d = 1 to M do
        Initialize γ_d
        while no local convergence do
            for p = 1 to N_d do
                for i = 1 to K do
                    Update φ_{d,p,i} using Equation (4.6)
                end for
                Normalize φ_{d,p,*} to sum to 1
            end for
            for i = 1 to K do
                Update γ_{d,i} using Equation (4.5)
            end for
        end while
    end for
    for i = 1 to K do
        for j = 1 to V do
            Update λ_{i,j} using Equation (4.4)
        end for
    end for
end while
```

4.4 Evaluation

We evaluate the model on two labeled datasets of tweets and one of titles of scientific publications. The evaluation targets multiple tasks such as topic coherence, document clustering and document classification. Please consult Chapter 2.3 for details about the metrics utilized to assess these tasks.

The following models are used in the evaluation for comparison:

- **Co-occurrence Topic Model (CTM)** This is the model proposed in this chapter. Common values for sparse priors are used ($\alpha = 0.1$; $\eta = 0.01$). We also use T=30 (i.e.: the number of additional co-occurring pairs we bring into the document for each existing pair; remember though that we use the set of the resulting collection to reduce noise and repetition). We found this setting to provide a good performance across the datasets. Nevertheless, we reiterate that identifying better ways to add related word co-occurrences to the documents is desirable. We leave this out to future work.

- **Latent Dirichlet Allocation (LDA)** We use the prior values recommended in previous work [36, 56] ($\alpha = 0.1$; $\eta = 0.01$). For a review of this model please consult Chapter 2.1.1.2.

- **Mixture of Unigrams (MoU)** We use the prior values recommended in previous work [56] ($\alpha = 50/K$; $\eta = 0.01$). For a review of this model please consult Chapter 2.1.1.1.

- **Biterm Topic Model (BTM)** This model proposed by Yan *et al.* [56] has been selected to represent the state of the art in short text data. The model has a preprocessing step in which all the biterms (i.e.: word pairs) of every document are generated. Then the biterms become the input of a Mixture of Unigrams model (i.e.: global topic proportions and one topic per biterm). For a fair comparison, we reimplemented the model with Variational Inference (original implementation is done using Gibbs Sampling). We use the same priors as in the original paper ($\alpha = 50/K$; $\eta = 0.01$).

- **SC-LDA FA** This is the SC-LDA model described in Chapter 3 with first author as context. This particular instance was the best performer from that chapter. We use it to assess whether utilizing context (when available) still leads to better results. Common values for sparse priors are used ($\alpha = 0.1$; $\eta = 0.01$).

The models are initialized according to standard practices from the literature. Blei & Lafferty [8] find that a good way to initialize the topics is to use a random sample of N documents from the corpus and compute a smoothed word distribution over the vocabulary space from the word counts of the random sample. We choose N to be 10.

We perform the evaluation with 3 levels of K (i.e.: the number of topics): $K = Z$, $K = 2Z$ and $K = 3Z$, where Z is the number of ground truth classes of a dataset. For each setting of a model we do 10 runs and report the result that has the maximum ELBO - the bigger the ELBO the closer the variational approximation is to the true posterior.

Dataset	Classes	Documents	Unique words	All words	Average words per document
FSD	21	2108	742	14457	6.858
Sanders	4	2073	1210	10366	5.000
arXiv	10	14587	2833	92518	6.342

Table 4.1: Statistics of the datasets used in the evaluation

4.4.1 Dataset Selection

We use in the evaluation three datasets of short text data. Two of the datasets contain tweets: First Story Detection (FSD) [38] and Sanders[1]. The other dataset consists of titles of scientific publications downloaded from arXiV.org, a dataset we also used in Chapter 3. Table 4.1 summarizes useful dataset statistics after preprocessing (basic stop and rare word removal). The ground truth classes of FSD correspond to events such as "Death of Amy Winehouse" or "Terrorist attack in Delhi"; the Sanders corpus contains hand classified tweets into 4 distinct categories (e.g.: "google", "microsoft"); while the arXiv dataset has labels which correspond to different areas from physics (e.g.: "Condensed Matter", "Nuclear Theory").

To facilitate the replicability of the experimental setup we note that the original FSD dataset contains 27 classes. We discarded 6 of them because they contain a small number of tweets (less than 10). The discarded classes are: "Topic 3: Betty Ford dies", "Topic 5: Flight Noar Linhas Aereas 4896 crashes, all 16 passengers dead", "Topic 11: Goran Hadzic, Yugoslavian war criminal, arrested", "Topic 12: India and Bangladesh sign a peace pact", "Topic 23: South Sudan becomes a UN member state", and "Topic 26: Rebels capture Tripoli international airport, Libya". In terms of preprocessing, on the Sanders and FSD datasets the following actions were taken: removed stop words, words with a length smaller than 3 characters, words with a global frequency smaller than 3 and discarded documents with less than 3 words. In addition to that, for the Sanders dataset we also had to remove the non-English tweets. For details about the arXiv dataset please consult Chapter 3.4.1 (where it is labeled as "Short Text Larger Contexts").

[1] Available at `http://www.sananalytics.com/lab/twitter-sentiment/`

Dataset	Method	Top 5 words	Top 10 words	Top 20 words
FSD	LDA	-8.834	-73.397	-434.231
	MoU	-10.354	-78.253	-436.299
	BMT	-7.513	-64.275	-409.479
	CTM	**-6.700**	**-58.522**	**-396.275**
Sanders	LDA	-14.723	-110.759	-552.352
	MoU	-13.666	-103.420	-552.043
	BTM	-14.076	-108.101	-545.227
	CTM	**-12.356**	**-94.394**	**-524.661**
arXiv	LDA	-21.803	-131.204	-680.870
	MoU	-19.035	-125.617	-627.604
	BTM	-18.206	-119.148	-606.940
	CTM	-17.771	**-116.640**	**-602.169**
	SC-LDA FA	**-17.628**	-118.817	-618.784

Table 4.2: Topic Coherence results with K set to the number of ground truth classes: K=21 for FSD, K=4 for Sanders and K=10 for arXiv.

4.4.2 Topic Coherence Evaluation

In this section we present and discuss the results for topic coherence, a measure of topic quality which aims to capture the human interpretability of topics in an automatic fashion (i.e.: no human annotators). Please consult Chapter 2.3.1 for a review of the task and details about the utilized metric.

In Table 4.2 we list the topic coherence results when K is set to the number of ground truth classes, whereas in Table 4.3 we show the trends as K varies. On the datasets that consist of tweets (i.e.: FSD and Sanders) CTM manages to clearly outperform all the other models across all levels of K (see both Tables 4.2 and 4.3). On the FSD dataset, the state of the art, BTM, comes second best, while on the Sanders dataset it falls behind MoU, occupying the third place. On the arXiv dataset, BTM and CTM tend to perform on par; MoU manages here to get the best results on the top 10 and 20 words for larger values of K (see Table 4.3). LDA is clearly the worst performing model on Sanders and arXiv datasets, reconfirming the negative effect sparsity has on this model (see Tables 4.2 and 4.3).

4.4.3 Document Clustering Evaluation

In this section we present and discuss the results for document clustering. We asses this task with three common metrics: Purity, Normalized Mutual Information and

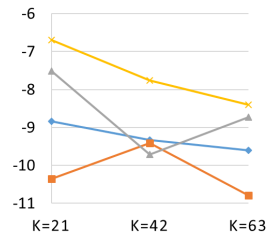
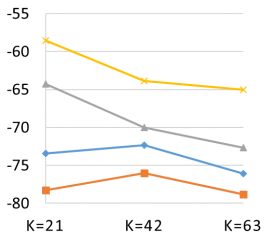
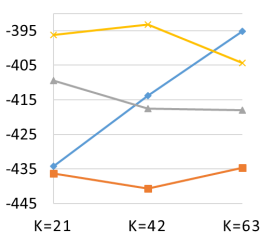
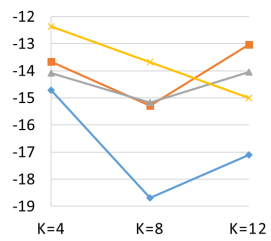
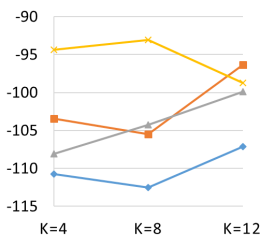
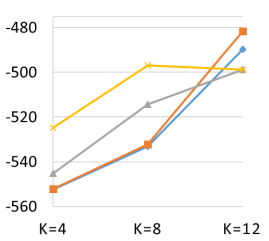
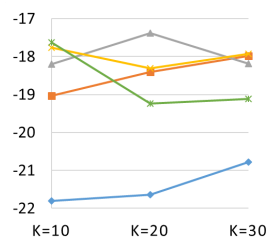
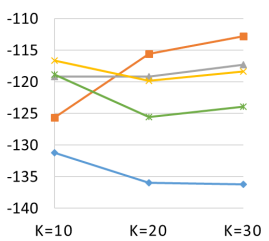
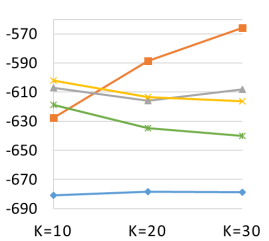
Dataset	Top 5 words	Top 10 words	Top 20 words
FSD			
Sanders			
arXiv			

Table 4.3: Topic Coherence results when K varies

Dataset	Method	Purity	NMI	ARI
FSD	LDA	0.857	0.801	0.726
	MoU	0.799	0.725	0.575
	BMT	0.894	0.815	0.597
	CTM	**0.915**	**0.849**	**0.604**
Sanders	LDA	0.481	0.108	0.097
	MoU	0.484	0.104	0.111
	BTM	0.523	**0.143**	**0.136**
	CTM	**0.537**	0.140	0.134
arXiv	LDA	0.578	0.233	0.195
	MoU	0.752	0.489	0.546
	BTM	0.805	0.533	0.448
	CTM	0.809	0.546	**0.607**
	SC-LDA FA	**0.820**	**0.584**	0.462

Table 4.4: Document Clustering results with K set to the number of ground truth classes: K=21 for FSD, K=4 for Sanders and K=10 for arXiv.

Adjusted Rand Index. Please consult Chapter 2.3.2 for a review of the task and details about the utilized metrics. Table 4.4 lists the results when K is set to the number of ground truth classes, while Table 4.5 shows the trends as K varies.

CTM manages to outperform, overall, the baselines (i.e.: MoU, LDA) and the state of the art (i.e.: BTM), on all the datasets, across all metrics, and levels of K (see Table 4.5). On the arXiv dataset, SC-LDA FA, with the help of contextual information, outperforms all the other models (see Purity and NMI from Table 4.5). LDA is overall the worst performer, while MoU comes second to last (best seen on the Sanders and arXiv datasets from Table 4.5). The state of the art, BTM, consistently ranks below our proposed model CTM, but on top of the baselines.

4.4.4 Document Classification Evaluation

In this section we present and discuss the results for document classification. We asses this task in terms of Accuracy. Please consult Chapter 2.3.3 for a review of the task and details about the utilized metric. Table 4.6 lists the results when K is set to the number of ground truth classes, while Table 4.7 shows the trends as K varies.

CTM outperforms all the other models on the datasets of tweets (see FSD and Sanders columns from Table 4.7). On the arXiv dataset, SC-LDA FA, with the

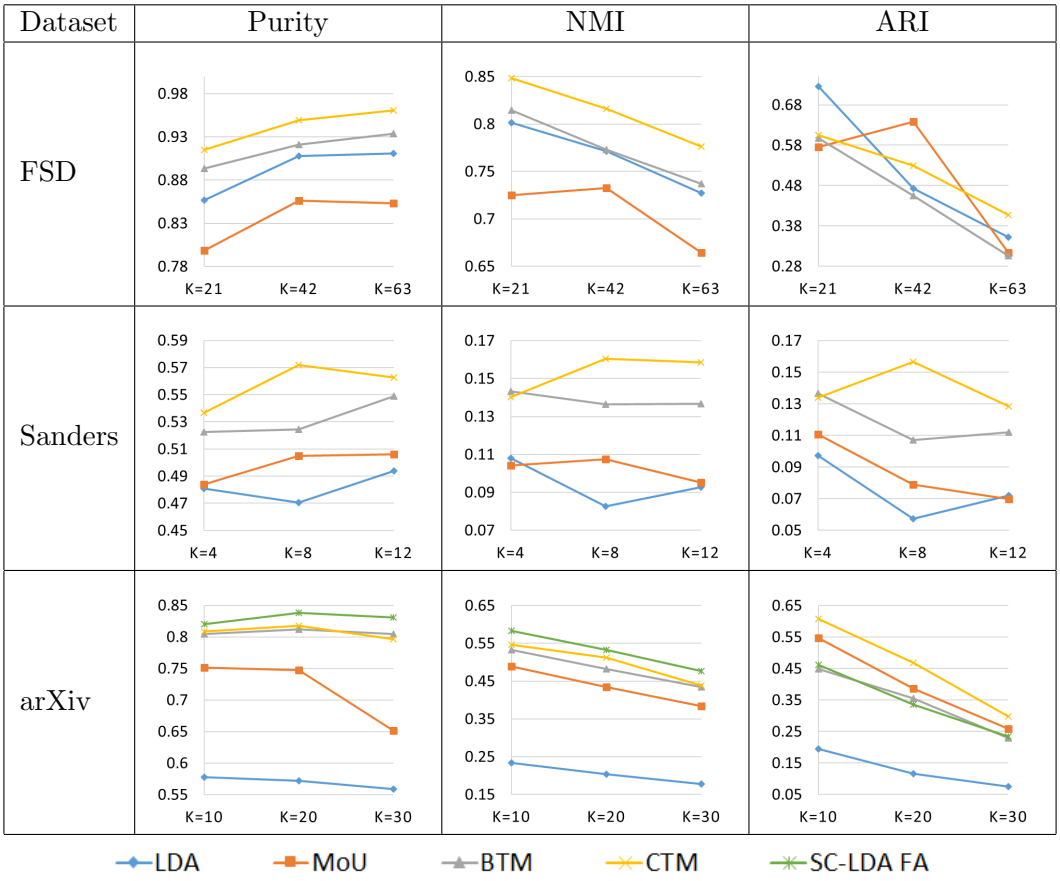


Table 4.5: Document Clustering results when K varies

Dataset	Method	Accuracy
FSD	LDA	0.873
	MoU	0.794
	BMT	0.917
	CTM	**0.940**
Sanders	LDA	0.498
	MoU	0.480
	BTM	**0.531**
	CTM	**0.530**
arXiv	LDA	0.583
	MoU	0.732
	BTM	0.818
	CTM	0.815
	SC-LDA FA	**0.840**

Table 4.6: Document Classification results with K set to the number of ground truth classes: K=21 for FSD, K=4 for Sanders and K=10 for arXiv.

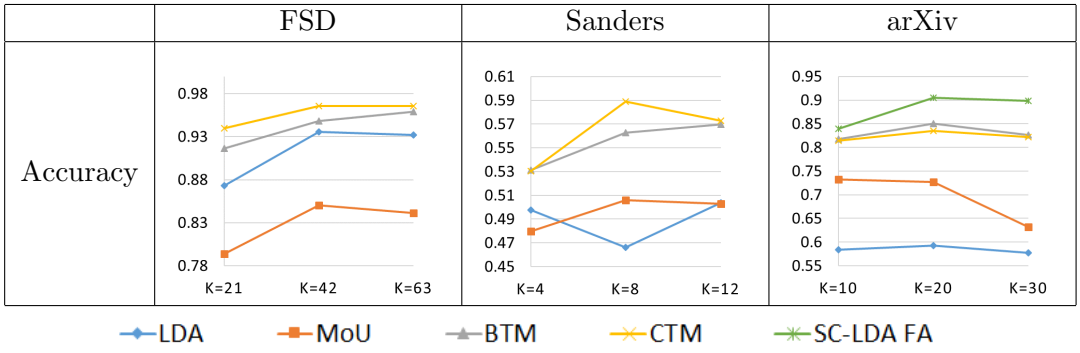


Table 4.7: Document Classification results when K varies

help of contextual information, has the best results across all levels of K (see last column of Table 4.7). On the same dataset, CTM and BTM perform on par. On the Sanders and arXiv datasets, the mixture model performs better than the admixture model (i.e.: MoU vs. LDA), the latter being drastically affected by sparsity (best seen on the arXiv dataset). This last result is more pronounced since MoU has the disadvantage brought by its "one topic per document" assumption which, in a classification task, produces only one feature to represent a document. The state of the art, BTM, consistently performs better than the baselines.

4.5 Discussion

As a follow-up to the discussion from Chapter 3 that topic models like LDA fail on short text data because of the lack of enough observations for a reliable inference, in this chapter we proposed a new topic model, which in contrast to our previous approach to alleviate sparsity, it can be used when contextual information is not available or it does not help. The introduced model was formulated around the observation that in normal text data, a classic model like LDA works well because patterns of word co-occurrences arise across the documents. In the generative process every document was modelled as a bag of word co-occurrences, where each co-occurrence belongs to a latent topic. The documents were enhanced a priori with related co-occurrences from the other documents, such that the collection had a greater chance of exhibiting word patterns.

We evaluated the model on two labeled datasets of tweets and one of titles of sci-

entific publications. The evaluation targeted multiple tasks such as topic coherence, document clustering and document classification. We list below the main findings:

- We find that, overall, our model surpasses the state of the art and the other baselines in terms of Topic Coherence, Document Clustering and Classification.
- The best performing contextual model from Chapter 3 (i.e.: SC-LDA FA) manages to get the best results in this evaluation as well in terms of Document Clustering and Classification, further strengthening the argument that contextual information is indeed useful when available.

Based on the assessments made in this chapter we can conclude that, overall, the proposed model brings an increase in performance when compared with the state of the art. We believe there is room for improvement, especially in the way related word co-occurrences are added to the documents, which we leave out to future work. It is also worth investigating the effect the enhancement has on the original topical representation of the documents. The added co-occurrences can lead to a concept drift (i.e.: an unforeseen change in the topical representations). Nevertheless, in our evaluation set up the model performed well. The results indicate that novel approaches which focus on modeling word co-occurrences are a promising direction towards a new class of models for short text data.

Chapter 5

Experimenting with a Subset Topic Model

Traditional topic models assume either a single topic per document, or a mixture of topics, where the number of mixture components is the same as the total number of topics the model aims to extract from the whole collection. However, neither of the aforementioned assumptions are entirely plausible. Even if the "one topic per document" assumption performs reasonably well on a short text dataset such as a Twitter collection, there can be many tweets which cover more than one topic. At the same time, even though longer documents tend to cover multiple topics, it is implausible they cover the whole topic space. In this chapter we experiment with a new topic model architecture which models documents using only a subset of the total number of topics. We compare the introduced model with the best known topic models that follow the aforementioned assumptions. The evaluation assesses coherence, a measure of topic interpretability, and is performed in varying text environments from very short to medium and longer text. The experiments indicate a connection between the size of the documents and the performance of the models with respect to the number of topics assumed for every document.

5.1 Motivation

Traditional topic model architectures such as Latent Dirichlet Allocation (LDA) and Mixture of Unigrams (MoU) aim to extract K topics from a data collection. While MoU assumes a document exhibits only one topic, at the other extreme, LDA models a document as a mixture of all the K topics. These distinct assumptions make the former more suitable in short text environments, and the latter a better fit in normal text data. [36, 56]. Yan *et al.* [56] mention in their work that LDA's poor performance in short text data is caused by the K-dimensional vector governing the per-document topic proportions - sparsity arises as its inference relies on a small number of observations (i.e.: the words in the document). In a study on the factors which affect the performance of LDA, Tang *et al.* [45] conclude that poor performance is expected when the documents are too short, even if you have a large collection. Another important conclusion of the study (with respect to the motivation behind this chapter) is that LDA is expected to perform better when the documents are associated with small subsets of topics.

In this chapter we argue that neither of the aforementioned assumptions of MoU and LDA are entirely plausible. Even if the "one topic per document" assumption performs reasonably well on a short text dataset such as a Twitter collection, there can be many tweets which cover more than one topic. At the same time, even though longer documents tend to cover multiple topics, it is implausible they cover the whole topic space. With these observations in mind, we propose a new topic model architecture which maintains the generic goal of discovering K topics in a corpus, but models documents as a mixture of only a subset of the topic space. The model aims to provide a generative process that is closer to a natural topical interpretation of the documents: if there are K topics in a corpus, then a document exhibits only a small subset of them.

In the evaluation we assess the performance of multiple instances of the model with different subset sizes. We also include MoU and LDA, the models with the extreme assumptions (one topic only vs. all topical space). The evaluation is performed in different text environments, covering very short, medium and longer text.

The characteristics of the datasets allow us to assess whether a smaller number of topics is sufficient in short text documents compared to collections of longer text where a slightly bigger number might be better suited. The evaluation targets topic coherence, a measure of topic quality which aims to capture the human interpretability of topics in an automatic fashion (i.e.: no human annotators).

5.2 Model Specification

In this section we describe a new topic model architecture based on subsets of topics. We will refer in our discussions to the proposed model as STM (Subset Topic Model).

The model takes as input a collection of documents indexed by $d \in \{1, 2, ..., M\}$. Every document d is a collection of words indexed by $n \in \{1, 2, ..., N_d\}$. The model uses a predefined collection of topic subsets indexed by $x \in \{1, 2, ..., S\}$. The length of each subset x is a fixed constant T. The elements of a subset are indexed by $p \in \{1, 2, ..., T\}$. The p's element of a subset x is one of the K topics - and can be accessed via the following operation $x.p = i$, where $i \in \{1, 2, ..., K\}$ is a topic.

The graphical model of STM is presented in Figure 5.1. The generative process is given below:

1. Draw proportions over the subsets $\pi \sim Dir_S(\delta)$
2. For every topic $i \in \{1, 2, ..., K\}$:
   (a) Draw a word distribution $\beta_i \sim Dir_V(\eta)$
3. For every document $d \in \{1, 2, ..., M\}$:
   (a) Draw a subset $t_d \sim Cat_S(\pi)$
   (b) Draw proportions over the indexes of a subset's elements $\theta_d \sim Dir_T(\alpha)$
   (c) For every word position $n \in \{1, 2, ..., N_d\}$:
      i. Draw an index of a subset's element $z_{d,n} \sim Cat_T(\theta_d)$
      ii. Draw word $w_{d,n} \sim Cat_V(\beta_{t_d.z_{d,n}})$

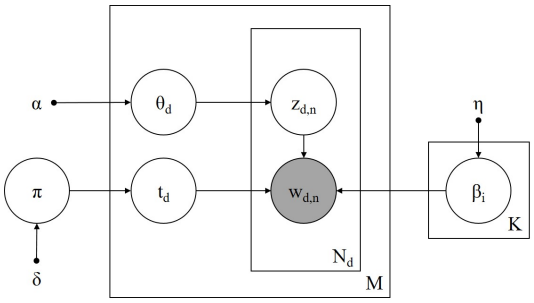


Figure 5.1: Graphical model of STM

5.2.1 Choosing the Subset Space

The model uses a predefined collection of topic subsets indexed by $x \in \{1, 2, ..., S\}$ where the length of each subset x is a fixed constant T. A natural choice for the subset space is to use all combinations of the K topics taken T at a time. In this case, $S = \binom{K}{T}$, can be too large for standard computation.

In this section we propose a way of generating a more manageable number S of subsets of length T formed with the K topics. Concretely, we are going to cut down some of the subsets generated by the combinations. We make the observation that in the space of combinations $S = \binom{K}{T}$ every topic i appears in the company of other topics $\binom{K-1}{T-1}$ times. We reduce S by constraining every topic i to appear in a smaller number of subsets. From a modeling perspective, this should not be a hard constraint, as it is unlikely the documents from a collection require a topic to appear in all the possible combinations with the other topics.

For completeness we introduce in Algorithm 3 the process by which all the combinations of the K topics taken T at a time are generated. At every level, you take into consideration all the smaller levels.

Algorithm 3 Generating combinations of K taken T

```
for i_1 = 1 to K do
    for i_2 = 1 to i_1 − 1 do
        ...
        for i_T = 1 to i_{T−1} − 1 do
            Generate (i_1, i_2, ..., i_T)
        end for
    end for
end for
```

Algorithm 4 outlines the process by which only some of the combinations of the

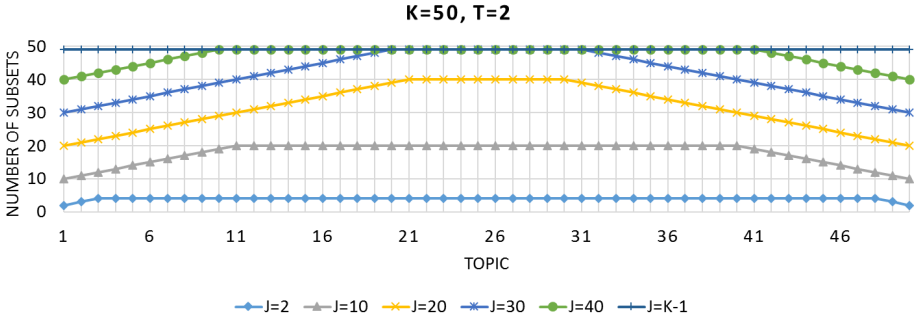


Figure 5.2: The number of subsets in which each topic appears as J varies.

K topics taken T at a time are generated. At every level, you take into consideration only the first J smaller levels, where J is a fixed constant. Note that when $J = K-1$ you generate the full space of combinations as in Algorithm 3.

Algorithm 4 Generating only some of the combinations of K taken T

1: **for** $i_1 = 1$ to K **do**
2: **for** $i_2 = \begin{cases} 1, & i_1 \leq J \\ i_1 - J, & i_1 > J \end{cases}$ to $i_1 - 1$ **do**
3: ...
4: **for** $i_T = \begin{cases} 1, & i_{T-1} \leq J \\ i_{T-1} - J, & i_{T-1} > J \end{cases}$ to $i_T - 1$ **do**
5: Generate $(i_1, i_2, ..., i_T)$
6: **end for**
7: **end for**
8: **end for**

The constant J controls the level of *approximation* with respect to the entire combinatorial space. Figure 5.2 illustrates an example of the behavior of Algorithm 4 when the combinatorial space is $\binom{K=50}{T=2}$.

In our experiments, we found the setting $J = 2$ to produce enough subsets for the STM model to perform well. Due to the lack of an obvious closed-form solution to the number of subsets produced by Algorithm 4, we list in Table 5.1 the number of subsets produced by the computer for different settings of K and T - this is to show we are dealing with a much smaller combinatorial space.

	T=2		**T=3**		**T=4**		**T=5**	
	J=2	J=K-1	**J=2**	J=K-1	**J=2**	J=K-1	**J=2**	J=K-1
K=10	17	45	28	120	44	210	64	252
K=20	37	190	68	1140	124	4845	224	15504
K=30	57	435	108	4060	204	27405	384	142506
K=40	77	780	148	9880	284	91390	544	658008
K=50	97	1225	188	19600	364	230300	704	2118760
K=60	117	1770	228	34220	444	487635	864	5461512
K=70	137	2415	268	54740	524	916895	1024	12103014
K=80	157	3160	308	82160	604	1581580	1184	24040016
K=90	177	4005	348	117480	684	2555190	1344	43949268
K=100	197	4950	388	161700	764	3921225	1504	75287520

Table 5.1: The number of subsets produced by Algorithm 4 with different values of K and T, keeping $J = 2$ fixed. For comparison purposes (with the entire combinatorial space), we also include the $J = K - 1$ setting.

5.3 Model Inference

To infer the latent parameters of the introduced model, we use standard variational inference, a deterministic technique for parameter estimation. Please consult Chapter 2.2.1 for a review. To keep things focused, we give here only an overview of the steps and derivations involved in the inference process - complementing material can be found in Appendix C.

The posterior of STM is presented in Equation (5.1) and factorizes according to the conditional dependencies from the graphical model presented in Figure 5.1. For simplicity we use symmetric priors on θ, β and π.

$$p(\theta, \beta, \pi, t, z|w, \alpha, \eta, \delta) \propto p(\theta|\alpha)p(\beta|\eta)p(\pi|\delta)p(t|\pi)p(z|\theta)p(w|z, t, \beta) \tag{5.1}$$

The variational distribution q used to approximate the STM posterior is presented in Equation (5.2) and factorizes according to the conditional dependencies from the graphical model presented in Figure 5.3.

$$q(\pi, \theta, \beta, t, z|\mu, \gamma, \lambda, \zeta, \phi) = q(\pi|\mu)q(\theta|\gamma)q(\beta|\lambda)q(t|\zeta)q(z|\phi) \tag{5.2}$$

With the posterior and the variational distribution at hand, we can define the

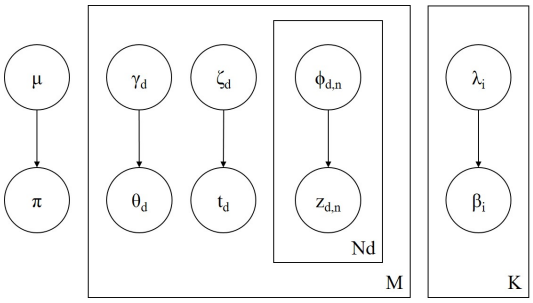


Figure 5.3: Graphical model of the variational distribution used to approximate the STM posterior

variational objective function. Equation (5.3) presents the ELBO in a compact form.

$$\begin{aligned}
\mathcal{L} &= E_q[\log p(\theta, \beta, \pi, w, z, t|\alpha, \eta, \delta)] - E_q[\log q(\theta, \beta, \pi, z, t|\gamma, \lambda, \mu, \phi, \zeta)] \\
&= E_q[\log p(\pi|\delta)] + E_q[\log p(\theta|\alpha)] + E_q[\log p(\beta|\eta)] + E_q[\log p(z|\theta)] + \\
&+ E_q[\log p(t|\pi)] + E_q[\log p(w|z, t, \beta)] - E_q[\log q(\theta|\gamma)] - \\
&- E_q[\log q(\beta|\lambda)] - E_q[\log q(\pi|\mu)] - E_q[\log q(z|\phi)] - E_q[\log q(t|\zeta)]
\end{aligned} \tag{5.3}$$

Maximizing the lower bound with respect to the variational parameters leads to the desired update formulas.

In Equation (5.4) we provide the update formula of the variational parameter associated with the subset proportions.

$$\mu_x = \delta_x + \sum_{d=1}^{M} \zeta_{d,x} \tag{5.4}$$

In Equation (5.5) we provide the update formula of the variational parameter associated with the document-level proportions over the indexes of a subset.

$$\gamma_{d,p} = \alpha_p + \sum_{n=1}^{N_d} \phi_{d,n,p} \tag{5.5}$$

In Equation (5.6) we provide the update formula of the variational parameter

associated with the subset assignment to a document.

$$\zeta_{d,x} \propto \exp\{\Psi(\mu_x) - \Psi(\mu_0) + \sum_{n,p,i,j}^{N_d,T,K,V} \phi_{d,n,p}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)\} \tag{5.6}$$

In Equation (5.7) we provide the update formula of the variational parameter associated with the assignment of an index of a subset's element to a word.

$$\phi_{d,n,p} \propto \exp\{\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0}) + \sum_{x,i,j}^{S,K,V} \zeta_{d,x}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)\} \tag{5.7}$$

In Equation (5.8) we provide the update formula of the variational parameter associated with a topic.

$$\lambda_{i,j} = \eta_j + \sum_{d,n,x,p}^{M,N_d,S,T} I(w_{d,n} = j)I(x.p = i)\zeta_{d,x}\phi_{d,n,p} \tag{5.8}$$

With the update formulas of the variational parameters at hand, the algorithm is straightforward. The variational parameters are updated iteratively until the lower bound from Equation (5.3) converges. This type of algorithm is known in the literature as Coordinate Ascent Mean-Field Variational Inference (CAVI) [10]. Algorithm 5 summarizes one iteration of CAVI.

5.3.1 Document-level Topic Proportions

STM does not model directly document-level topic proportions like, for example, LDA does. In this section we present a formula to generate this information using the estimated variational parameters.

Equation (5.9) gives the probability of topic i in document d. The formula is intuitive: the mass topic i receives in document d is based on the mass in document d of the subsets which contain topic i and the proportion of the topic in the document. Summing Equation (5.9) over i from 1 to K will give a result of one - an easy proof to show that we have indeed a probability distribution representing the document-level

Algorithm 5 One iteration of Mean Field Variational Inference for STM.

```
for d = 1 to M do
    for n = 1 to N_d do
        for p = 1 to T do
            Update φ_{d,n,p} using Equation (5.7)
        end for
        Normalize φ_{d,n,*} to sum to 1
    end for
    for x = 1 to S do
        Update ζ_{d,x} using Equation (5.6)
    end for
    Normalize ζ_{d,*} to sum to 1
    for p = 1 to T do
        Update γ_{d,p} using Equation (5.5)
    end for
end for
for x = 1 to S do
    Update μ_x using Equation (5.4)
end for
for i = 1 to K do
    for j = 1 to V do
        Update λ_{i,j} using Equation (5.8)
    end for
end for
```

topic proportions.

$$p(topic = i|d) = \sum_{x=1}^{S}(\zeta_{d,x} \sum_{p=1}^{T} I(x.p = i) \frac{\gamma_{d,p}}{\sum_{p=1}^{T} \gamma_{d,p}}) \tag{5.9}$$

5.4 Evaluation

The evaluation aims to assess the effect of constraining the number of topics a document can exhibit on the performance of the model given the characteristics of the input collection. The evaluation is done with respect to topic coherence, a measure of topic quality, in datasets showcasing different text environments, from very short, to medium and longer text.

The following models are used in the evaluation for comparison:

- **Latent Dirichlet Allocation (LDA)** This is a model in which documents are mixtures of K topics (the size of the entire topic space). We use the prior values recommended in previous work [36, 56] ($\alpha = 0.1$; $\eta = 0.01$). For a

review of this model please consult Chapter 2.1.1.2.

- **Mixture of Unigrams (MoU)** This is a model in which documents exhibit only one topic. We use the prior values recommended in previous work [56] ($\alpha = 50/K$; $\eta = 0.01$). For a review of this model please consult Chapter 2.1.1.1.

- **Subset Topic Model (STM)** We evaluate four instances of STM in which we fix the size of the subsets (number of maximum topics per document) to constants ranging from $T = 2$ to $T = 5$. The range of the per-document topics is selected to be plausible for short, medium and longer text collections. We use the approximation technique for the combinatorial space described in Section 5.2.1 with J=2. In terms of prior selection, we use common sparse priors for the word distribution of a topic ($\eta = 0.01$) and for the proportions over the subsets ($\delta = 0.1$). For the per-document proportions over the indexes of a subset's elements we use a uniform, non-informative prior ($\alpha = 1.0$).

The models are initialized according to standard practices from the literature. Blei & Lafferty [8] find that a good way to initialize the topics is to use a random sample of N documents from the corpus and compute a smoothed word distribution over the vocabulary space from the word counts of the random sample. We choose N to be 10.

We perform the evaluation with 3 levels of K (i.e.: number of topics): $K = Z$, $K = 2Z$ and $K = 3Z$, where Z is the number of ground truth classes of a dataset. For each setting of a model we do 10 runs and report the result that has the maximum ELBO - the bigger the ELBO the closer the variational approximation is to the true posterior.

5.4.1 Dataset Selection

The evaluation is performed on four datasets, covering multiple text environments from very short, to medium and longer text. The datasets used are 20 Newsgroup (20NG) [11], Reuters 8 (R8) [11], Tag My News (TMN) [48] and a titles-only version of Tag My News (TMN-T). Table 5.2 summarizes useful dataset statistics after

Dataset	Classes	Documents	Unique words	All words	Average words per document
20 Newsgroup (20NG)	20	18780	9899	1798945	95.79
Reuters 8 (R8)	8	9863	5252	511255	51.84
Tag My News (TMN)	7	32600	8621	557101	17.09
Titles of Tag My News (TMN-T)	7	30130	6303	152689	5.07

Table 5.2: Statistics of the datasets used in the evaluation

preprocessing (basic stop and rare word removal). It is worth pointing out the last column from the table, which indicates the wide spectrum of text environments. The 20NG dataset contains 20 ground truth classes which correspond to a variety of topics from Computer Graphics, to Motorcycles, Baseball and Religion. The Reuters dataset has categories like Earn, Grain, Trade, or Interest. The TMN and TMN-T datasets contain documents from 7 generic categories covering Sport, Business, U.S., Health, Sci&Tech, World and Entertainment. The number of documents per ground truth class is relatively balanced in the 20NG dataset, and more sparse in the others.

To facilitate the replicability of the experimental setup we discuss the details behind the preprocessing. For all the datasets, we discarded stop words, words with a length smaller than 3 characters and documents with less than 3 words. For the R8 and TMN datasets we discarded the words with a global frequency less than 10. For the TMN-T dataset the frequency threshold was 5, while for 20NG it was 30.

5.4.2 Topic Coherence Evaluation

In this section we present and discuss the results for topic coherence, a measure of topic quality which aims to capture the human interpretability of topics in an automatic fashion (i.e.: no human annotators). Please consult Chapter 2.3.1 for a review of the task and details about the utilized metric.

We focus first on Table 5.3. The results indicate that using a model which allows K topics per document (i.e.: LDA) leads to the worst coherence scores compared with all the other models, across all columns, on very short to short text (TMN-T and TMN datasets). At the other extreme, using a model which allows only one

Dataset	Method	Top 5 words	Top 10 words	Top 20 words
TMN-T	MoU	**-22.778**	**-146.090**	**-738.455**
	LDA	-29.395	-165.145	-778.575
	STM T2 J2	-25.398	-155.147	-745.603
	STM T3 J2	-24.834	-157.862	-773.100
	STM T4 J2	-25.573	-161.482	-776.503
	STM T5 J2	-25.063	-159.160	-773.580
TMN	MoU	-19.895	-124.986	-616.045
	LDA	-20.850	-123.513	-622.296
	STM T2 J2	-17.092	**-110.920**	**-580.669**
	STM T3 J2	-19.416	-117.573	-594.668
	STM T4 J2	**-17.009**	-115.882	-594.538
	STM T5 J2	-17.992	-116.690	-606.951
R8	MoU	-10.774	-76.348	-410.018
	LDA	-9.253	-71.768	-350.030
	STM T2 J2	-8.849	-69.822	-380.413
	STM T3 J2	-6.961	-57.728	-357.312
	STM T4 J2	**-6.349**	**-46.942**	**-307.390**
	STM T5 J2	-7.184	-48.305	-322.277
20NG	MoU	-11.616	**-73.009**	-394.146
	LDA	**-10.317**	-74.227	-380.699
	STM T2 J2	-10.972	-73.515	-393.910
	STM T3 J2	-10.872	-75.059	**-370.687**
	STM T4 J2	-12.249	-75.634	-390.196
	STM T5 J2	-11.524	-74.971	-389.671

Table 5.3: Topic Coherence results with K set to the number of ground truth classes: $K = 7$ for TMN-T and TMN; $K = 8$ for R8; $K = 20$ for 20NG.

topic per document (MoU), leads to poor performance on medium to longer text (R8 and 20NG datasets). On very short text data (TMN-T dataset) the model with one topic per document (i.e.: MoU) does best - increasing the number of topics per document (STM T2 J2 to STM T5 J2) is mainly inversely proportional to the performance (best seen in Top 10 and 20 words columns), suggesting that a lower number of topics gives a better performance. Moving on from the TMN-T dataset to the TMN dataset we have a difference in the average words per document of 12 (statistic taken from Table 5.2). The difference in the number of words causes the performance of MoU to drop significantly (best seen in Top 10 and 20 words columns) - this suggests that more than one topic is now required; instances of STM with $T = 2$ to $T = 5$ confirm that by having the better performance. Going from the TMN dataset to the R8 dataset, the number of average words per document triples. The one topic per-document assumption of MoU causes to model to become the worst performer. LDA is now better than MoU but its assumption of a maximum of K topics per document is still too broad for the length of the documents, placing its performance behind the one obtained by the instances of STM. On the 20NG dataset, many models perform on par when coherence is assessed on the top 5 and 10 words. A more clear advantage is achieved by STM T3 on Top 20 words.

In Figure 5.4, the evaluation set-up from Table 5.3 is replicated on different levels of K. On the very short text dataset (TMN-T), LDA keeps the previously identified pattern as the worst performer across all K values. On the TMN dataset, where documents are slightly lengthier (17 words on average), you have the same pattern as before in which LDA and MoU are the worst performers. Moving on to the R8 dataset, where the number of words triples (51 words on average), the results are mixed, but there are still some useful observations. For example, MoU is the worst performer overall on coherence scores on the top 10 and 20 words. STM T4 and T5 are in most cases better than LDA. This last pattern is also kept on the 20NG dataset.

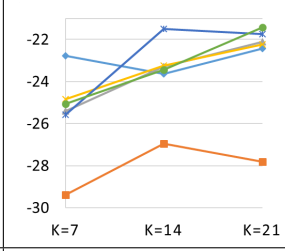
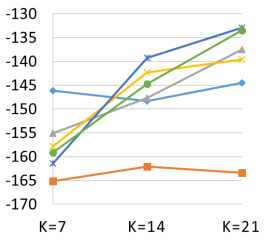
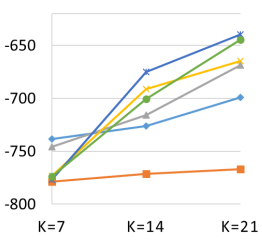
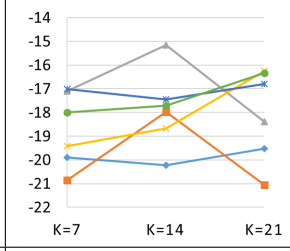
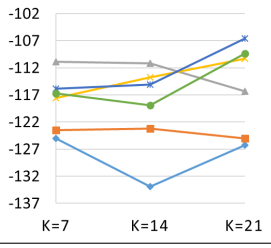
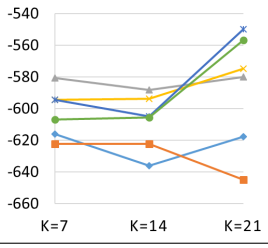
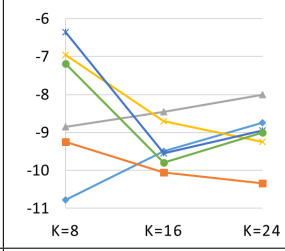
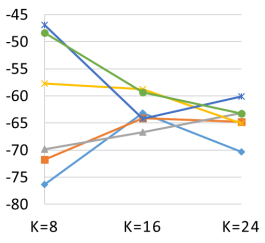
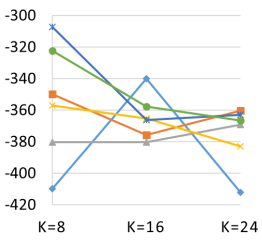
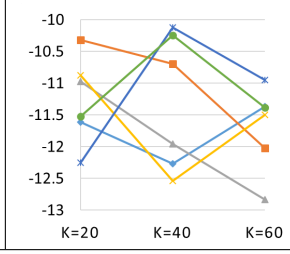
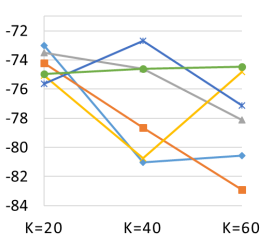
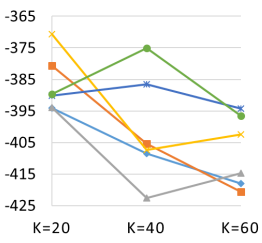
Dataset	Top 5 words	Top 10 words	Top 20 words
TMN-T			
TMN			
R8			
20NG			

Table 5.4: Topic Coherence results when K varies

5.5 Discussion

In this chapter we experimented with a topic model which assumes documents exhibit only a subset of the entire topic space. This complements existing work which assumes documents contain either a single topic - MoU - or a mixture of the entire topic space - LDA.

On very short text items (i.e.: 5 words on average) the evaluation confirms the previously known superiority of MoU and the drastic impact sparsity has on LDA. In this chapter we find though that on a dataset of items which are slightly bigger (i.e.: 17 words on average), instances of the proposed model with two to five topics per document perform better than both MoU and LDA. On the longer text datasets (i.e.: 51 and 95 words on average per documents), we find that LDA is surpassed by the models with subset sizes of 4 and 5.

The proposed subset topic model has its drawbacks. Mainly, a better way of creating the subsets of topics and assigning them to the documents in desirable (i.e.: having a generative process to guide the assignment of latent topics into latent subsets). Future work can look into utilizing a Markov chain for this purpose (e.g.: creating the subset by adding topics conditioned to the ones already present; doing so, the subsets can be viewed as clusters of correlated topics). Even though the evaluation is sometimes noisy, the results indicate a connection between the size of the documents and the performance of the models with respect to the number of topics assumed for every document.

Chapter 6

Conclusions

Topic models have been used with great success over the years in organizing large collections of unstructured text allowing people to interact with the data more easily. The applicability is extensive and in multiple areas: analyzing the evolution of topics over time in digital library data [7, 54], the identification of correlated topics [4, 25], modeling authors and their publications [40], or capturing spatial and temporal patterns from blog posts [32, 33]. The advancements in parameter estimation techniques allow topic models to be applied at scale and in online frameworks [19, 20].

Although there is a vast research literature on topic models, the development of such models for short text data is still a relatively new field. Topic models which behave well on normal text collections under-perform on short text items due to a reduced number of observations (i.e.: the words) available for a reliable inference. This causes the models to suffer from sparsity.

In this dissertation this sparsity problem was addressed from two main perspectives. In the first part, we developed models which exploit the context that accompanies certain short text collections. Concretely, we utilized the authors in datasets created from titles of scientific publications, but other useful examples of context include hashtags for twitter data, locations for titles of blog posts or time for headlines of news articles. In the second part, we proposed a more general purpose model which can be used when such contextual information is not available. The model creates and exploits patterns of word co-occurrences. The evaluation

addressed multiple tasks such as topic coherence (i.e.: a measure of topic quality which aims to capture the human interpretability of topics in an automatic fashion), document clustering and document classification. We discuss below, in more details, the main contributions that result from this thesis with direct references to the posed research questions and the chapters which addressed them.

Which class of models benefits more from aggregation in short text data, a mixture or an admixture? Can document aggregation lead to state of the art performance?

In Chapter 3 we explicitly modeled the implicit assumptions of document aggregation, a popular heuristic employed to alleviate sparsity, and applied it to two standard model architectures: a mixture and an admixture. The latter is known to suffer greatly from sparsity, whereas the "one topic assumption" of the former is considered to be a good fit for short text items. The two architectures are also the backbone of a great number of models developed over the years. For evaluation, we created datasets with both very short (i.e.: titles of publications) and medium (i.e.: abstracts) text items, which also had different opportunities for aggregation (a smaller vs. a larger number of documents per context). This allowed us to assess the performance of the models with respect to different text environments and context sizes. Our findings indicate that an admixture model benefits more from aggregation compared to a mixture which rarely improved over its baseline (i.e.: the standard mixture). We also find that the state of the art in short text data can be surpassed as long as every context contains a small number of documents. The findings inform future researchers interested in developing topic models for context accompanied short text data that having at the core of the models the set of assumptions of an admixture has the potential to lead to a better performance compared to developing a model on top of a mixture.

Can short text collections be enhanced such that repeating word co-occurrences have a better chance to arise across the documents more consistently and facilitate a better topic discovery?

In Chapter 4 we introduced a new topic model, which in contrast to our previous approach to alleviate sparsity, can be used when contextual information is not available or it does not help (i.e.: it is shared by documents which have little or no

topical relationship). The model proposed was formulated around the observation that in normal text data, a classic model like LDA works well because patterns of word co-occurrences arise across the documents. However, the possibility of such patterns to arise in a short text dataset is reduced. The model assumes every document is a bag of word co-occurrences, where each co-occurrence belongs to a latent topic. The documents were enhanced a priori with related co-occurrences from the other documents, such that the collection had a greater chance of exhibiting word patterns. We evaluated the model on two labeled datasets of tweets and one of titles of scientific publications. The latter is a dataset which we also utilized in Chapter 3 and has contextual information available. The model we proposed performed well managing to surpass the state of the art and popular topic model baselines. The best performing contextual model introduced in Chapter 3 managed to get the best results in this evaluation as well, further strengthening the argument that contextual information is indeed useful when available. Nevertheless, the results showed that novel approaches which focus on modeling word co-occurrences are a promising direction towards a new class of models for short text data.

Can topic models be improved by assuming a more appropriate number of topics for every document?

In Chapter 5 we experimented with a topic model which assumes documents are mixtures of only a subset of the entire topic space. This complements existing work which assumes documents contain either a single topic or a mixture of the entire topic space. The model was built on the observation that the aforementioned assumptions are too extreme. Even if the "one topic per document" assumption performs reasonably well on a short text dataset such as a Twitter collection, there can be many tweets which cover more than one topic. At the same time, even though longer documents tend to cover multiple topics, it is implausible they cover the whole topic space. The evaluation assessed coherence, a measure of topic interpretability, and was performed in varying text environments from very short to medium and longer text. The results, although preliminary, were in accordance with the observations made and indicated a connection between the size of the documents and the performance of the models with respect to the number of topics assumed

for every document. The findings from this chapter inform researchers that topic models trained on short text data could obtain a better performance not only by increasing the number of observations, but also by reducing the size of the topic space associated with the documents.

Appendix A

Detailed Proofs for Single-Context Topic Models

Throughout the proofs we make use of a short-hand notation which de-clutters the mathematics. For some K-dimensional vector α, we use the convention $\alpha_0 = \sum_{i=1}^{K} \alpha_i$. We also make a note of the digamma function $\Psi()$ present in many equations - this is the first derivative of the $\log \Gamma$ function and can be computed using a Taylor approximation [1].

A.1 The Single-Context Mixture of Unigrams Model

Equations (A.1), (A.2), (A.3) and (A.4) complete the description of the model from Chapter 3.2.2.

Equation (A.1) represents the probability of the context specific topic proportions in the exponential family form.

$$p(\theta_x|\alpha) = \exp\{(\sum_{i=1}^{K}(\alpha_i - 1)\log\theta_{x,i}) + \log\Gamma(\alpha_0) - \sum_{i=1}^{K}\log\Gamma(\alpha_i)\} \tag{A.1}$$

Equation (A.2) represents the probability of a topic in the exponential family form.

$$p(\beta_i|\eta) = \exp\{(\sum_{j=1}^{V}(\eta_j - 1)\log\beta_{i,j}) + \log\Gamma(\eta_0) - \sum_{j=1}^{V}\log\Gamma(\eta_j)\} \tag{A.2}$$

Equation (A.3) gives the probability of a topic assignment for a document.

$$p(z_d|\theta) = \theta_{c_d,z_d} = \prod_{x,i}^{C,K} \theta_{x,i}^{I(c_d=x)I(z_d=i)} \tag{A.3}$$

Equation (A.4) gives the probability of a word given the topic assigned to the document it belongs to.

$$p(w_{d,n}|z_d,\beta) = \beta_{z_d,w_{d,n}} = \prod_{i,j}^{K,V} \beta_{i,j}^{I(w_{d,n}=j)I(z_d=i)} \tag{A.4}$$

A.1.1 Deriving the Complete Conditionals

In this section we derive the complete conditionals of every latent variable given all the other latent variables and the observations. We are showing that each such conditional is in the exponential family. We further define the variational distributions to have the same form as their corresponding complete conditionals.

In Equation (A.5) we derive the complete conditional associated with the context topic proportions.

$$\begin{aligned}
p(\theta_x|\theta_-, z, \beta, w) &\propto p(\theta_x|\alpha)p(z|\theta_x) \\
&\propto p(\theta_x|\alpha) \prod_{d=1}^{M} p(z_d|\theta_x)^{I(c_d=x)} \\
&\propto \prod_{i=1}^{K} \theta_{x,i}^{\alpha_i - 1} \prod_{d,i}^{M,K} \theta_{x,i}^{I(c_d=x)I(z_d=i)} \\
&\propto \prod_{i=1}^{K} \theta_{x,i}^{[\alpha_i + \sum_{d=1}^{M} I(c_d=x)I(z_d=i)]-1} \\
&= Dir(a), a_i = \alpha_i + \sum_{d=1}^{M} I(c_d = x)I(z_d = i) \\
&= \exp\{(\sum_{i=1}^{K}(a_i - 1)\log\theta_{x,i}) + \log\Gamma(a_0) - \sum_{i=1}^{K}\log\Gamma(a_i)\}
\end{aligned} \tag{A.5}$$

Because the complete conditional of the context topic proportions is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well.

Equations (A.6) and (A.7) give the necessary information.

$$q(\theta_x|\gamma_x) = \exp\{(\sum_{i=1}^{K}(\gamma_{x,i}-1)\log\theta_{x,i}) + \log\Gamma(\gamma_{x,0}) - \sum_{i=1}^{K}\log\Gamma(\gamma_{x,i})\} \tag{A.6}$$

$$E_q[\log\theta_{x,i}|\gamma_x] = \Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0}) \tag{A.7}$$

In Equation (A.8) we derive the complete conditional associated with the topics.

$$\begin{aligned}
p(\beta_i|\beta_-, z, \theta, w) &\propto p(\beta_i|\eta)p(w|z,\beta_i) \\
&\propto p(\beta_i|\eta)\prod_{d,n}^{M,N_d} p(w_{d,n}|z_d,\beta_i) \\
&\propto \prod_{j=1}^{V}\beta_{i,j}^{\eta_j-1}\prod_{d,n,j}^{M,N_d,V}\beta_{i,j}^{I(w_{d,n}=j)I(z_d=i)} \\
&\propto \prod_{j=1}^{V}\beta_{i,j}^{[\eta_j+\sum_{d,n}^{M,N_d}I(w_{d,n}=j)I(z_d=i)]-1} \\
&= Dir(b), b_j = \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n}=j)I(z_d=i) \\
&= \exp\{(\sum_{j=1}^{V}(b_j-1)\log\beta_{i,j}) + \log\Gamma(b_0) - \sum_{j=1}^{V}\log\Gamma(b_j)\}
\end{aligned} \tag{A.8}$$

Because the complete conditional of a topic is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well. Equations (A.9) and (A.10) give the necessary information.

$$q(\beta_i|\lambda_i) = \exp\{(\sum_{j=1}^{V}(\lambda_{i,j}-1)\log\beta_{i,j}) + \log\Gamma(\lambda_{i,0}) - \sum_{j=1}^{V}\log\Gamma(\lambda_{i,j})\} \tag{A.9}$$

$$E_q[\log\beta_{i,j}|\lambda i] = \Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}) \tag{A.10}$$

In Equation (A.11) we derive the complete conditional associated with the per-

document topic assignments.

$$
\begin{aligned}
p(z_d = i|z_-, \theta, \beta, w) &\propto p(z_d = i|\theta)p(w|\beta_i) \\
&\propto p(z_d = i|\theta) \prod_{n=1}^{N_d} p(w_{d,n}|\beta_i) \\
&\propto \prod_{x=1}^{C} \theta_{x,i}^{I(c_d=x)} \prod_{n,j}^{N_d,V} \beta_{i,j}^{I(w_{d,n}=j)} \\
&\propto \theta_{c_d,i} \prod_{n=1}^{N_d} \beta_{i,w_{d,n}} \\
&\propto \exp\{\log c\}, c = \theta_{c_d,i} \prod_{n=1}^{N_d} \beta_{i,w_{d,n}} \\
&\propto \exp\{\log \theta_{c_d,i} + \sum_{n=1}^{N_d} \log \beta_{i,w_{d,n}}\}
\end{aligned}
\tag{A.11}
$$

Because the complete conditional of the per-document topic assignment is a Categorical, the corresponding variational distribution is going to be a Categorical as well. Equation (A.12) gives the necessary information.

$$
q(z_d = i|\phi_d) = \phi_{d,i} = \exp\{\log \phi_{d,i}\} \tag{A.12}
$$

A.1.2 Deriving the Update Formulas of the Variational Parameters

The mathematics of the inference are based on the fact that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

In Equation (A.13) we derive the update formula of the variational parameter

associated with the topics.

$$\begin{aligned}
\lambda_{i,j} - 1 &= E_q[b_j - 1] \\
\lambda_{i,j} = E_q[b_j] &= E_q[\eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j) I(z_d = i)] \\
&= \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j) E_q[I(z_d = i)] \\
&= \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j) \phi_{d,i}
\end{aligned} \tag{A.13}$$

In Equation (A.14) we derive the update formula of the variational parameter associated with the context topic proportions.

$$\begin{aligned}
\gamma_{x,i} - 1 &= E_q[a_i - 1] \\
\gamma_{x,i} = E_q[a_i] &= E_q[\alpha_i + \sum_{d=1}^{M} I(c_d = x) I(z_d = i)] \\
&= \alpha_i + \sum_{d=1}^{M} I(c_d = x) E_q[I(z_d = i)] \\
&= \alpha_i + \sum_{d=1}^{M} I(c_d = x) \phi_{d,i}
\end{aligned} \tag{A.14}$$

In Equation (A.15) we derive the update formula of the variational parameter associated with the per-document topic assignments.

$$\begin{aligned}
\log \phi_{d,i} \propto E_q[\log c] &= E_q[\log(\theta_{c_d,i} \prod_{n=1}^{N_d} \beta_{i,w_{d,n}})] \\
&= E_q[\log(\prod_{x=1}^{C} \theta_{x,i}^{I(c_d=x)} \prod_{n,j}^{N_d,V} \beta_{i,j}^{I(w_{d,n}=j)})] \\
&= \sum_{x=1}^{C} I(c_d = x) E_q[\log \theta_{x,i}] + \sum_{n,j}^{N_d,V} I(w_{d,n} = j) E_q[\log \beta_{i,j}] \\
&= \sum_{x=1}^{C} I(c_d = i)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{n,j}^{N_d,V} I(w_{d,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) \\
\phi_{d,i} \propto \exp\{\sum_{x=1}^{C} I(c_d = x)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) &+ \sum_{n,j}^{N_d,V} I(w_{d,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\}
\end{aligned} \tag{A.15}$$

A.1.3 Deriving the Evidence Lower Bound

The ELBO is the objective function which needs to be maximized. The maximization is done using a coordinate ascent algorithm in which the variational parameters are updated iteratively until the ELBO convergences. Monitoring the value of the ELBO is useful for assessing algorithm termination, but also for sanity checks (the ELBO is guaranteed to increase with every iteration).

In Equation (A.16) we expand the lower bound according to the conditional dependencies of the model and those of the variational distribution.

$$\begin{aligned}\mathcal{L} &= E_q[\log p(\theta, \beta, w, z|\alpha, \eta)] - E_q[\log q(\theta, \beta, z|\gamma, \lambda, \phi)] \\ &= E_q[\log p(\theta|\alpha)] + E_q[\log p(\beta|\eta)] + E_q[\log p(z|\theta)] + E_q[\log p(w|z, \beta)] \\ &- E_q[\log q(\theta|\gamma)] - E_q[\log q(\beta|\lambda)] - E_q[\log q(z|\phi)]\end{aligned} \tag{A.16}$$

In Equation (A.17) we derive the expectation term that regards the probability of the topic proportions.

$$\begin{aligned}E_q[\log p(\theta|\alpha)] &= E_q[\log \prod_{x=1}^{C} p(\theta_x|\alpha)] \\ &= E_q[\sum_{x=1}^{C} \log p(\theta_x|\alpha)] \\ &= \sum_{x,i}^{C,K} (\alpha_i - 1)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{x=1}^{C} \log \Gamma(\alpha_0) - \sum_{x,i}^{C,K} \log \Gamma(\alpha_i)\end{aligned} \tag{A.17}$$

In Equation (A.18) we derive the expectation term that regards the probability of the topics.

$$\begin{aligned}E_q[\log p(\beta|\eta)] &= E_q[\log \prod_{i=1}^{K} p(\beta_i|\eta)] \\ &= E_q[\sum_{i=1}^{K} \log p(\beta_i|\eta)] \\ &= \sum_{i,j}^{K,V} (\eta_j - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\eta_0) - \sum_{i,j}^{K,V} \log \Gamma(\eta_j)\end{aligned} \tag{A.18}$$

In Equation (A.19) we derive the expectation term that regards the probability of the topic assignments.

$$\begin{aligned} E_q[\log p(z|\theta)] &= E_q[\log \prod_{d=1}^{M} p(z_d|\theta)] \\ &= E_q[\log \prod_{d=1}^{M} \theta_{c_d,z_d}] \\ &= E_q[\log \prod_{d,x,i}^{M,C,K} \theta_{x,i}^{I(c_d=x)I(z_d=i)}] \\ &= \sum_{d,x,i}^{M,C,K} I(c_d = x)\phi_{d,i}(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) \end{aligned} \tag{A.19}$$

In Equation (A.20) we derive the expectation term that regards the probability of the words.

$$\begin{aligned} E_q[\log p(w|z,\beta)] &= E_q[\log \prod_{d,n}^{M,N_d} p(w_{d,n}|z_d,\beta)] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \beta_{z_d,w_{d,n}}] \\ &= E_q[\log \prod_{d,n,i,j}^{M,N_d,K,V} \beta_{i,j}^{I(w_{d,n}=j)I(z_d=i)}] \\ &= \sum_{d,n,i,j}^{M,N_d,K,V} I(w_{d,n} = j)\phi_{d,i}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) \end{aligned} \tag{A.20}$$

In Equation (A.21) we derive the expectation term that regards the variational distributions of the topic proportions.

$$\begin{aligned} E_q[\log q(\theta|\gamma)] &= E_q[\log \prod_{x=1}^{C} q(\theta_x|\gamma_x)] \\ &= E_q[\sum_{x=1}^{C} \log q(\theta_x|\gamma_x)] \\ &= \sum_{x,i}^{C,K} (\gamma_{x,i} - 1)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{x=1}^{C} \log\Gamma(\gamma_{x,0}) - \sum_{x,i}^{C,K} \log\Gamma(\gamma_{x,i}) \end{aligned} \tag{A.21}$$

In Equation (A.22) we derive the expectation term that regards the variational

distributions of the topics.

$$\begin{aligned} E_q[\log q(\beta|\lambda)] &= E_q[\log \prod_{i=1}^{K} p(\beta_i|\lambda_i)] \\ &= E_q[\sum_{i=1}^{K} \log q(\beta_i|\lambda_i)] \\ &= \sum_{i,j}^{K,V} (\lambda_{i,j} - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\lambda_{i,0}) - \sum_{i,j}^{K,V} \log \Gamma(\lambda_{i,j}) \end{aligned} \tag{A.22}$$

In Equation (A.23) we derive the expectation term that regards the variational distributions of the topic assignments.

$$\begin{aligned} E_q[\log q(z|\phi)] &= E_q[\log \prod_{d=1}^{M} q(z_d|\phi_d)] \\ &= E_q[\log \prod_{d=1}^{M} \phi_{d,z_d}] \\ &= E_q[\log \prod_{d,i}^{M,K} \phi_{d,i}^{I(z_d=i)}] \\ &= \sum_{d,i}^{M,K} \phi_{d,i} \log \phi_{d,i} \end{aligned} \tag{A.23}$$

A.2 The Single-Context Latent Dirichlet Allocation Model

Equations (A.24), (A.25), (A.26) and (A.27) complete the description of the model from Chapter 3.2.1.

Equation (A.24) represents the probability of the context specific topic proportions in the exponential family form.

$$p(\theta_x|\alpha) = \exp\{(\sum_{i=1}^{K} (\alpha_i - 1) \log \theta_{x,i}) + \log \Gamma(\alpha_0) - \sum_{i=1}^{K} \log \Gamma(\alpha_i)\} \tag{A.24}$$

Equation (A.25) represents the probability of a topic in the exponential family

form.

$$p(\beta_i|\eta) = \exp\{(\sum_{j=1}^{V}(\eta_j - 1)\log\beta_{i,j}) + \log\Gamma(\eta_0) - \sum_{j=1}^{V}\log\Gamma(\eta_j)\} \tag{A.25}$$

Equation (A.26) gives the probability of a topic assignment for a word.

$$p(z_{d,n}|\theta) = \theta_{c_d,z_{d,n}} = \prod_{x,i}^{C,K}\theta_{x,i}^{I(c_d=x)I(z_{d,n}=i)} \tag{A.26}$$

Equation (A.27) gives the probability of a word given its assigned topic.

$$p(w_{d,n}|z_{d,n},\beta) = \beta_{z_{d,n},w_{d,n}} = \prod_{i,j}^{K,V}\beta_{i,j}^{I(w_{d,n}=j)I(z_{d,n}=i)} \tag{A.27}$$

A.2.1 Deriving the Complete Conditionals

In this section we derive the complete conditionals of every latent variable given all the other latent variables and the observations. We are showing that each such conditional is in the exponential family. We further define the variational distributions to have the same form as their corresponding complete conditionals.

In Equation (A.28) we derive the complete conditional associated with the per-context topic proportions.

$$\begin{aligned}
p(\theta_x|\theta_-, z, \beta, w) &\propto p(\theta_x|\alpha)p(z|\theta_x) \\
&\propto p(\theta_x|\alpha)\prod_{d,n}^{M,N_d} p(z_{d,n}|\theta_x)^{I(c_d=x)} \\
&\propto \prod_{i=1}^{K}\theta_{x,i}^{\alpha_i-1}\prod_{d,n,i}^{M,N_d,K}\theta_{x,i}^{I(c_d=x)I(z_{d,n}=i)} \\
&\propto \prod_{i=1}^{K}\theta_{x,i}^{[\alpha_i+\sum_{d,n}^{M,N_d}I(c_d=x)I(z_{d,n}=i)]-1} \\
&= Dir(a), a_i = \alpha_i + \sum_{d,n}^{M,N_d} I(c_d = x)I(z_{d,n} = i) \\
&= \exp\{(\sum_{i=1}^{K}(a_i - 1)\log\theta_{x,i}) + \log\Gamma(a_0) - \sum_{i=1}^{K}\log\Gamma(a_i)\}
\end{aligned} \tag{A.28}$$

Because the complete conditional of the per-context topic proportions is a Dirich-

let, the corresponding variational distribution is going to be a Dirichlet as well. Equations (A.29) and (A.30) give the necessary information.

$$q(\theta_x|\gamma_x) = \exp\{(\sum_{i=1}^{K}(\gamma_{x,i}-1)\log\theta_{x,i}) + \log\Gamma(\gamma_{x,0}) - \sum_{i=1}^{K}\log\Gamma(\gamma_{x,i})\} \tag{A.29}$$

$$E_q[\log\theta_{x,i}|\gamma_x] = \Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0}) \tag{A.30}$$

In Equation (A.31) we derive the complete conditional associated with the topics.

$$\begin{aligned}
p(\beta_i|\beta_-, z, \theta, w) &\propto p(\beta_i|\eta)p(w|z,\beta_i) \\
&\propto p(\beta_i|\eta)\prod_{d,n}^{M,N_d} p(w_{d,n}|z_{d,n},\beta_i) \\
&\propto \prod_{j=1}^{V}\beta_{i,j}^{\eta_j-1}\prod_{d,n,j}^{M,N_d,V}\beta_{i,j}^{I(w_{d,n}=j)I(z_{d,n}=i)} \\
&\propto \prod_{j=1}^{V}\beta_{i,j}^{[\eta_j+\sum_{d,n}^{M,N_d}I(w_{d,n}=j)I(z_{d,n}=i)]-1} \\
&= Dir(b), b_j = \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n}=j)I(z_{d,n}=i) \\
&= \exp\{(\sum_{j=1}^{V}(b_j-1)\log\beta_{i,j}) + \log\Gamma(b_0) - \sum_{j=1}^{V}\log\Gamma(b_j)\}
\end{aligned} \tag{A.31}$$

Because the complete conditional of a topic is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well. Equations (A.32) and (A.33) give the necessary information.

$$q(\beta_i|\lambda_i) = \exp\{(\sum_{j=1}^{V}(\lambda_{i,j}-1)\log\beta_{i,j}) + \log\Gamma(\lambda_{i,0}) - \sum_{j=1}^{V}\log\Gamma(\lambda_{i,j})\} \tag{A.32}$$

$$E_q[\log\beta_{i,j}|\lambda i] = \Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}) \tag{A.33}$$

In Equation (A.34) we derive the complete conditional associated with the per

word topic assignments.

$$\begin{aligned} p(z_{d,n} = i|z_-, \theta, \beta, w) &\propto p(z_{d,n} = i|\theta)p(w_{d,n}|\beta_i) \\ &\propto \prod_{x=1}^{C} \theta_{x,i}^{I(c_d=x)} \prod_{j=1}^{V} \beta_{i,j}^{I(w_{d,n}=j)} \\ &\propto \theta_{c_d,i}\beta_{i,w_{d,n}} \\ &\propto \exp\{\log c\}, c = \theta_{c_d,i}\beta_{i,w_{d,n}} \\ &\propto \exp\{\log \theta_{c_d,i} + \log \beta_{i,w_{d,n}}\} \end{aligned} \tag{A.34}$$

Because the complete conditional of the per-word topic assignment is a Categorical, the corresponding variational distribution is going to be a Categorical as well. Equation (A.35) gives the necessary information.

$$q(z_{d,n} = i|\phi_{d,n}) = \phi_{d,n,i} = \exp\{\log \phi_{d,n,i}\} \tag{A.35}$$

A.2.2 Deriving the Update Formulas of the Variational Parameters

The mathematics of the inference are based on the fact that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

In Equation (A.36) we derive the update formula of the variational parameter associated with the topics.

$$\begin{aligned} \lambda_{i,j} - 1 &= E_q[b_j - 1] \\ \lambda_{i,j} &= E_q[b_j] = E_q[\eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)I(z_{d,n} = i)] \\ &= \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)E_q[I(z_{d,n} = i)] \\ &= \eta_j + \sum_{d,n}^{M,N_d} I(w_{d,n} = j)\phi_{d,n,i} \end{aligned} \tag{A.36}$$

In Equation (A.37) we derive the update formula of the variational parameter

associated with the per-context topic proportions.

$$\begin{aligned}
\gamma_{x,i} - 1 &= E_q[a_i - 1] \\
\gamma_{x,i} = E_q[a_i] &= E_q[\alpha_i + \sum_{d,n}^{M,N_d} I(c_d = x) I(z_{d,n} = i)] \\
&= \alpha_i + \sum_{d,n}^{M,N_d} I(c_d = x) E_q[I(z_{d,n} = i)] \\
&= \alpha_i + \sum_{d,n}^{M,N_d} I(c_d = x) \phi_{d,n,i}
\end{aligned} \tag{A.37}$$

In Equation (A.38) we derive the update formula of the variational parameter associated with the per-document topic assignments.

$$\begin{aligned}
\log \phi_{d,n,i} \propto E_q[\log c] &= E_q[\log(\theta_{c_d,i} \beta_{i,w_{d,n}})] \\
&= E_q[\log(\prod_{x=1}^{C} \theta_{x,i}^{I(c_d=x)} \prod_{j=1}^{V} \beta_{i,j}^{I(w_{d,n}=j)})] \\
&= \sum_{x=1}^{C} I(c_d = x) E_q[\log \theta_{x,i}] + \sum_{j=1}^{V} I(w_{d,n} = j) E_q[\log \beta_{i,j}] \\
&= \sum_{x=1}^{C} I(c_d = x)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{j=1}^{V} I(w_{d,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) \\
\phi_{d,n,i} \propto \exp\{&\sum_{x=1}^{C} I(c_d = x)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{j=1}^{V} I(w_{d,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\}
\end{aligned} \tag{A.38}$$

A.2.3 Deriving the Evidence Lower Bound

The ELBO is the objective function which needs to be maximized. The maximization is done using a coordinate ascent algorithm in which the variational parameters are updated iteratively until the ELBO convergences. Monitoring the value of the ELBO is useful for assessing algorithm termination, but also for sanity checks (the ELBO is guaranteed to increase with every iteration).

In Equation (A.39) we expand the lower bound according to the conditional

dependencies of the model and those of the variational distribution.

$$\begin{aligned}\mathcal{L} &= E_q[\log p(\theta, \beta, w, z|\alpha, \eta)] - E_q[\log q(\theta, \beta, z|\gamma, \lambda, \phi)] \\ &= E_q[\log p(\theta|\alpha)] + E_q[\log p(\beta|\eta)] + E_q[\log p(z|\theta)] + E_q[\log p(w|z, \beta)] \\ &- E_q[\log q(\theta|\gamma)] - E_q[\log q(\beta|\lambda)] - E_q[\log q(z|\phi)]\end{aligned} \tag{A.39}$$

In Equation (A.40) we derive the expectation term that regards the probability of the topic proportions.

$$\begin{aligned}E_q[\log p(\theta|\alpha)] &= E_q[\log \prod_{x=1}^{C} p(\theta_x|\alpha)] \\ &= E_q[\sum_{x=1}^{C} \log p(\theta_x|\alpha)] \\ &= \sum_{x,i}^{C,K} (\alpha_i - 1)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{x=1}^{C} \log \Gamma(\alpha_0) - \sum_{x,i}^{C,K} \log \Gamma(\alpha_i)\end{aligned} \tag{A.40}$$

In Equation (A.41) we derive the expectation term that regards the probability of the topics.

$$\begin{aligned}E_q[\log p(\beta|\eta)] &= E_q[\log \prod_{i=1}^{K} p(\beta_i|\eta)] \\ &= E_q[\sum_{i=1}^{K} \log p(\beta_i|\eta)] \\ &= \sum_{i,j}^{K,V} (\eta_j - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\eta_0) - \sum_{i,j}^{K,V} \log \Gamma(\eta_j)\end{aligned} \tag{A.41}$$

In Equation (A.42) we derive the expectation term that regards the probability

of the topic assignments.

$$\begin{aligned} E_q[\log p(z|\theta)] &= E_q[\log \prod_{d,n}^{M,N_d} p(z_{d,n}|\theta)] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \theta_{c_d,z_{d,n}}] \\ &= E_q[\log \prod_{d,n,x,i}^{M,N_d,C,K} \theta_{x,i}^{I(c_d=x)I(z_{d,n}=i)}] \\ &= \sum_{d,n,x,i}^{M,N_d,C,K} I(c_d = x)\phi_{d,n,i}(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) \end{aligned} \tag{A.42}$$

In Equation (A.43) we derive the expectation term that regards the probability of the words.

$$\begin{aligned} E_q[\log p(w|z,\beta)] &= E_q[\log \prod_{d,n}^{M,N_d} p(w_{d,n}|z_{d,n},\beta)] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \beta_{z_{d,n},w_{d,n}}] \\ &= E_q[\log \prod_{d,n,i,j}^{M,N_d,K,V} \beta_{i,j}^{I(w_{d,n}=j)I(z_{d,n}=i)}] \\ &= \sum_{d,n,i,j}^{M,N_d,K,V} I(w_{d,n} = j)\phi_{d,n,i}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) \end{aligned} \tag{A.43}$$

In Equation (A.44) we derive the expectation term that regards the variational distributions of the topic proportions.

$$\begin{aligned} E_q[\log q(\theta|\gamma)] &= E_q[\log \prod_{x=1}^{C} q(\theta_x|\gamma_x)] \\ &= E_q[\sum_{x=1}^{C} \log q(\theta_x|\gamma_x)] \\ &= \sum_{x,i}^{C,K} (\gamma_{x,i} - 1)(\Psi(\gamma_{x,i}) - \Psi(\gamma_{x,0})) + \sum_{x=1}^{C} \log\Gamma(\gamma_{x,0}) - \sum_{x,i}^{C,K} \log\Gamma(\gamma_{x,i}) \end{aligned} \tag{A.44}$$

In Equation (A.45) we derive the expectation term that regards the variational

distributions of the topics.

$$\begin{aligned} E_q[\log q(\beta|\lambda)] &= E_q[\log \prod_{i=1}^{K} q(\beta_i|\lambda_i)] \\ &= E_q[\sum_{i=1}^{K} \log q(\beta_i|\lambda_i)] \\ &= \sum_{i,j}^{K,V} (\lambda_{i,j} - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\lambda_{i,0}) - \sum_{i,j}^{K,V} \log \Gamma(\lambda_{i,j}) \end{aligned} \tag{A.45}$$

In Equation (A.46) we derive the expectation term that regards the variational distributions of the topic assignments.

$$\begin{aligned} E_q[\log q(z|\phi)] &= E_q[\log \prod_{d,n}^{M,N_d} q(z_{d,n}|\phi_{d,n})] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \phi_{d,n,z_{d,n}}] \\ &= E_q[\log \prod_{d,n,i}^{M,N_d,K} \phi_{d,n,i}^{I(z_{d,n}=i)}] \\ &= \sum_{d,n,i}^{M,N_d,K} \phi_{d,n,i} \log \phi_{d,n,i} \end{aligned} \tag{A.46}$$

A. DETAILED PROOFS FOR SINGLE-CONTEXT TOPIC MODELS

Appendix B

Detailed Proofs for the Co-occurrence Topic Model

Throughout the proof, we make use of a short-hand notation which de-clutters the mathematics. For some K-dimensional vector α, we use the convention $\alpha_0 = \sum_{i=1}^{K} \alpha_i$. We also make a note of the digamma function $\Psi()$ present in many equations - this is the first derivative of the $\log \Gamma$ function and can be computed using a Taylor approximation [1].

Equations (B.1), (B.2), (B.3) and (B.4) complete the description of the model from Chapter 4.2.

Equation (B.1) represents the probability of the document-level topic proportions in the exponential family form.

$$p(\theta_d|\alpha) = \exp\{(\sum_{i=1}^{K}(\alpha_i - 1)\log\theta_{d,i}) + \log\Gamma(\alpha_0) - \sum_{i=1}^{K}\log\Gamma(\alpha_i)\} \tag{B.1}$$

Equation (B.2) represents the probability of a topic in the exponential family form.

$$p(\beta_i|\eta) = \exp\{(\sum_{j=1}^{V}(\eta_j - 1)\log\beta_{i,j}) + \log\Gamma(\eta_0) - \sum_{j=1}^{V}\log\Gamma(\eta_j)\} \tag{B.2}$$

Equation (B.3) gives the probability of a topic assignment to a word co-occurrence.

$$p(z_{d,p}|\theta_d) = \theta_{d,z_{d,p}} = \prod_{i=1}^{K} \theta_{d,i}^{I(z_{d,p}=i)} \tag{B.3}$$

Equation (B.4) gives the probability of a word given the topic assigned to the co-occurrence set it is part of.

$$p(w_{d,p,n}|z_{d,p}, \beta) = \beta_{z_{d,p},w_{d,p,n}} = \prod_{i,j}^{K,V} \beta_{i,j}^{I(w_{d,p,n}=j)I(z_{d,p}=i)} \tag{B.4}$$

B.1 Deriving the Complete Conditionals

In this section we derive the complete conditionals of every latent variable given all the other latent variables and the observations. We are showing that each such conditional is in the exponential family. We further define the variational distributions to have the same form as their corresponding complete conditionals.

In Equation (B.5) we derive the complete conditional associated with the document topic proportions.

$$\begin{aligned}
p(\theta_d|\theta_-, z, \beta, w) &\propto p(\theta_d|\alpha)p(z|\theta_d) \\
&\propto p(\theta_d|\alpha) \prod_{p=1}^{N_d} p(z_{d,p}|\theta_d) \\
&\propto \prod_{i=1}^{K} \theta_{d,i}^{\alpha_i - 1} \prod_{p,i}^{N_d,K} \theta_{d,i}^{I(z_{d,p}=i)} \\
&\propto \prod_{i=1}^{K} \theta_{d,i}^{[\alpha_i + \sum_{p=1}^{N_d} I(z_{d,p}=i)]-1} \\
&= Dir(a), a_i = \alpha_i + \sum_{p=1}^{N_d} I(z_{d,p} = i) \\
&= \exp\{(\sum_{i=1}^{K} (a_i - 1) \log \theta_{d,i}) + \log \Gamma(a_0) - \sum_{i=1}^{K} \log \Gamma(a_i)\}
\end{aligned} \tag{B.5}$$

Because the complete conditional of the document topic proportions is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well.

Equations (B.6) and (B.7) give the necessary information.

$$q(\theta_d|\gamma_d) = \exp\{(\sum_{i=1}^{K}(\gamma_{d,i}-1)\log\theta_{d,i}) + \log\Gamma(\gamma_{d,0}) - \sum_{i=1}^{K}\log\Gamma(\gamma_{d,i})\} \tag{B.6}$$

$$E_q[\log\theta_{d,i}|\gamma_d] = \Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0}) \tag{B.7}$$

In Equation (B.8) we derive the complete conditional associated with the topics.

$$\begin{aligned}
p(\beta_i|\beta_-, z, \theta, w) &\propto p(\beta_i|\eta)p(w|z,\beta_i) \\
&\propto p(\beta_i|\eta)\prod_{d,p,n}^{M,N_d,N_{d,p}} p(w_{d,p,n}|z_{d,p},\beta_i) \\
&\propto \prod_{j=1}^{V}\beta_{i,j}^{\eta_j-1}\prod_{d,n,p,j}^{M,N_d,N_{d,p},V}\beta_{i,j}^{I(w_{d,p,n}=j)I(z_{d,p}=i)} \\
&\propto \prod_{j=1}^{V}\beta_{i,j}^{[\eta_j+\sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n}=j)I(z_{d,p}=i)]-1} \\
&= Dir(b), b_j = \eta_j + \sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n}=j)I(z_{d,p}=i) \\
&= \exp\{(\sum_{j=1}^{V}(b_j-1)\log\beta_{i,j}) + \log\Gamma(b_0) - \sum_{j=1}^{V}\log\Gamma(b_j)\}
\end{aligned} \tag{B.8}$$

Because the complete conditional of a topic is a Dirichlet, the corresponding variational distribution is going to be a Dirichlet as well. Equations (B.9) and (B.10) give the necessary information.

$$q(\beta_i|\lambda_i) = \exp\{(\sum_{j=1}^{V}(\lambda_{i,j}-1)\log\beta_{i,j}) + \log\Gamma(\lambda_{i,0}) - \sum_{j=1}^{V}\log\Gamma(\lambda_{i,j})\} \tag{B.9}$$

$$E_q[\log\beta_{i,j}|\lambda i] = \Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}) \tag{B.10}$$

In Equation (B.11) we derive the complete conditional associated with the per

word co-occurrence topic assignments.

$$\begin{aligned} p(z_{d,p} = i|z_-, \theta, \beta, w) &\propto p(z_{d,p} = i|\theta_d)p(w|\beta_i) \\ &\propto p(z_{d,p} = i|\theta_d) \prod_{n=1}^{N_{d,p}} p(w_{d,p,n}|\beta_i) \\ &\propto \theta_{d,i} \prod_{n,j}^{N_{d,p},V} \beta_{i,j}^{I(w_{d,p,n}=j)} \\ &\propto \theta_{d,i} \prod_{n=1}^{N_{d,p}} \beta_{i,w_{d,p,n}} \\ &\propto \exp\{\log c\}, c = \theta_{d,i} \prod_{n=1}^{N_{d,p}} \beta_{i,w_{d,p,n}} \\ &\propto \exp\{\log \theta_{d,i} + \sum_{n=1}^{N_{d,p}} \log \beta_{i,w_{d,p,n}}\} \end{aligned} \tag{B.11}$$

Because the complete conditional of the per word co-occurrence topic assignment is a Categorical, the corresponding variational distribution is going to be a Categorical as well. Equation (B.12) gives the necessary information.

$$q(z_{d,p} = i|\phi_{d,p}) = \phi_{d,p,i} = \exp\{\log \phi_{d,p,i}\} \tag{B.12}$$

B.2 Deriving the Update Formulas of the Variational Parameters

The mathematics of the inference are based on the fact that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

In Equation (B.13) we derive the update formula of the variational parameter

associated with the topics.

$$\begin{aligned}
\lambda_{i,j} - 1 &= E_q[b_j - 1] \\
\lambda_{i,j} = E_q[b_j] &= E_q[\eta_j + \sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n} = j) I(z_{d,p} = i)] \\
&= \eta_j + \sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n} = j) E_q[I(z_{d,p} = i)] \\
&= \eta_j + \sum_{d,p,n}^{M,N_d,N_{d,p}} I(w_{d,p,n} = j) \phi_{d,p,i}
\end{aligned} \tag{B.13}$$

In Equation (B.14) we derive the update formula of the variational parameter associated with the document topic proportions.

$$\begin{aligned}
\gamma_{d,i} - 1 &= E_q[a_i - 1] \\
\gamma_{d,i} = E_q[a_i] &= E_q[\alpha_i + \sum_{p=1}^{N_d} I(z_{d,p} = i)] \\
&= \alpha_i + \sum_{p=1}^{N_d} E_q[I(z_{d,p} = i)] \\
&= \alpha_i + \sum_{p=1}^{N_d} \phi_{d,p,i}
\end{aligned} \tag{B.14}$$

In Equation (B.15) we derive the update formula of the variational parameter associated with the per word co-occurrence topic assignments.

$$\begin{aligned}
\log \phi_{d,p,i} \propto E_q[\log c] &= E_q[\log(\theta_{d,i} \prod_{n=1}^{N_{d,p}} \beta_{i,w_{d,p,n}})] \\
&= E_q[\log(\theta_{d,i} \prod_{n,j}^{N_{d,p},V} \beta_{i,j}^{I(w_{d,p,n}=j)})] \\
&= E_q[\log \theta_{d,i}] + \sum_{n,j}^{N_{d,p},V} I(w_{d,p,n} = j) E_q[\log \beta_{i,j}] \\
&= \Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0}) + \sum_{n,j}^{N_{d,p},V} I(w_{d,p,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) \\
\phi_{d,p,i} \propto \exp\{\Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0}) &+ \sum_{n,j}^{N_{d,p},V} I(w_{d,p,n} = j)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\}
\end{aligned} \tag{B.15}$$

B.3 Deriving the Evidence Lower Bound

The ELBO is the objective function which needs to be maximized. The maximization is done using a coordinate ascent algorithm in which the variational parameters are updated iteratively until the ELBO convergences. Monitoring the value of the ELBO is useful for assessing algorithm termination, but also for sanity checks (the ELBO is guaranteed to increase with every iteration).

In Equation (B.16) we expand the lower bound according to the conditional dependencies of the model and those of the variational distribution.

$$\begin{aligned}\mathcal{L} &= E_q[\log p(\theta, \beta, w, z|\alpha, \eta)] - E_q[\log q(\theta, \beta, z|\gamma, \lambda, \phi)] \\ &= E_q[\log p(\theta|\alpha)] + E_q[\log p(\beta|\eta)] + E_q[\log p(z|\theta)] + E_q[\log p(w|z, \beta)] \\ &- E_q[\log q(\theta|\gamma)] - E_q[\log q(\beta|\lambda)] - E_q[\log q(z|\phi)]\end{aligned} \tag{B.16}$$

In Equation (B.17) we derive the expectation term that regards the probability of the topic proportions.

$$\begin{aligned}E_q[\log p(\theta|\alpha)] &= E_q[\log \prod_{d=1}^{M} p(\theta_d|\alpha)] \\ &= E_q[\sum_{d=1}^{M} \log p(\theta_d|\alpha)] \\ &= \sum_{d,i}^{M,K} (\alpha_i - 1)(\Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0})) + \sum_{d=1}^{M} \log \Gamma(\alpha_0) - \sum_{d,i}^{M,K} \log \Gamma(\alpha_i)\end{aligned} \tag{B.17}$$

In Equation (B.18) we derive the expectation term that regards the probability

of the topics.

$$\begin{aligned} E_q[\log p(\beta|\eta)] &= E_q[\log \prod_{i=1}^{K} p(\beta_i|\eta)] \\ &= E_q[\sum_{i=1}^{K} \log p(\beta_i|\eta)] \\ &= \sum_{i,j}^{K,V} (\eta_j - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\eta_0) - \sum_{i,j}^{K,V} \log \Gamma(\eta_j) \end{aligned} \tag{B.18}$$

In Equation (B.19) we derive the expectation term that regards the probability of the topic assignments.

$$\begin{aligned} E_q[\log p(z|\theta)] &= E_q[\log \prod_{d,p}^{M,N_d} p(z_{d,p}|\theta_d)] \\ &= E_q[\log \prod_{d,p}^{M,N_d} \theta_{d,z_{d,p}}] \\ &= E_q[\log \prod_{d,p,i}^{M,N_d,K} \theta_{d,i}^{I(z_{d,p}=i)}] \\ &= \sum_{d,p,i}^{M,N_d,K} \phi_{d,p,i}(\Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0})) \end{aligned} \tag{B.19}$$

In Equation (B.20) we derive the expectation term that regards the probability of the words.

$$\begin{aligned} E_q[\log p(w|z,\beta)] &= E_q[\log \prod_{d,p,n}^{M,N_d,N_{d,p}} p(w_{d,p,n}|z_{d,p},\beta)] \\ &= E_q[\log \prod_{d,p,n}^{M,N_d,N_{d,p}} \beta_{z_{d,p},w_{d,p,n}}] \\ &= E_q[\log \prod_{d,p,n,i,j}^{M,N_d,N_{d,p},K,V} \beta_{i,j}^{I(w_{d,p,n}=j)I(z_{d,p}=i)}] \\ &= \sum_{d,p,n,i,j}^{M,N_d,N_{d,p},K,V} I(w_{d,p,n}=j)\phi_{d,p,i}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) \end{aligned} \tag{B.20}$$

In Equation (B.21) we derive the expectation term that regards the variational

distributions of the topic proportions.

$$\begin{aligned} E_q[\log q(\theta|\gamma)] &= E_q[\log \prod_{d=1}^{M} q(\theta_d|\gamma_d)] \\ &= E_q[\sum_{d=1}^{M} \log q(\theta_d|\gamma_d)] \\ &= \sum_{d,i}^{M,K} (\gamma_{d,i} - 1)(\Psi(\gamma_{d,i}) - \Psi(\gamma_{d,0})) + \sum_{d=1}^{M} \log \Gamma(\gamma_{d,0}) - \sum_{d,i}^{M,K} \log \Gamma(\gamma_{d,i}) \end{aligned} \tag{B.21}$$

In Equation (B.22) we derive the expectation term that regards the variational distributions of the topics.

$$\begin{aligned} E_q[\log q(\beta|\lambda)] &= E_q[\log \prod_{i=1}^{K} p(\beta_i|\lambda_i)] \\ &= E_q[\sum_{i=1}^{K} \log q(\beta_i|\lambda_i)] \\ &= \sum_{i,j}^{K,V} (\lambda_{i,j} - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\lambda_{i,0}) - \sum_{i,j}^{K,V} \log \Gamma(\lambda_{i,j}) \end{aligned} \tag{B.22}$$

In Equation (B.23) we derive the expectation term that regards the variational distributions of the topic assignments.

$$\begin{aligned} E_q[\log q(z|\phi)] &= E_q[\log \prod_{d,p}^{M,N_d} q(z_{d,p}|\phi_{d,p})] \\ &= E_q[\log \prod_{d,p}^{M,N_d} \phi_{d,p,z_{d,p}}] \\ &= E_q[\log \prod_{d,p,i}^{M,N_d,K} \phi_{d,p,i}^{I(z_{d,p}=i)}] \\ &= \sum_{d,p,i}^{M,N_d,K} \phi_{d,p,i} \log \phi_{d,p,i} \end{aligned} \tag{B.23}$$

Appendix C

Detailed Proofs for the Subset Topic Model

Throughout the proof, we make use of a short-hand notation which de-clutters the mathematics. For some K-dimensional vector α, we use the convention $\alpha_0 = \sum_{i=1}^{K} \alpha_i$. We also make a note of the digamma function $\Psi()$ present in many equations - this is the first derivative of the $\log \Gamma$ function and can be computed using a Taylor approximation [1].

Equations (C.1), (C.2), (C.3), (C.4), (C.5) and (C.6) complete the description of the model from Chapter 5.2.

Equation (C.1) represents the probability of the subset proportions in the exponential family form.

$$p(\pi|\delta) = \exp\{(\sum_{x=1}^{S}(\delta_x - 1)\log \pi_x) + \log\Gamma(\delta_0) - \sum_{x=1}^{S}\log\Gamma(\delta_x)\} \tag{C.1}$$

Equation (C.2) represents the probability of a topic in the exponential family form.

$$p(\beta_i|\eta) = \exp\{(\sum_{j=1}^{V}(\eta_j - 1)\log \beta_{i,j}) + \log\Gamma(\eta_0) - \sum_{j=1}^{V}\log\Gamma(\eta_j)\} \tag{C.2}$$

Equation (C.3) represents the probability of the proportions over the indexes of

a subset's elements in the exponential family form.

$$p(\theta_d|\alpha) = \exp\{(\sum_{p=1}^{T}(\alpha_p - 1)\log\theta_{d,p}) + \log\Gamma(\alpha_0) - \sum_{p=1}^{T}\log\Gamma(\alpha_p)\} \tag{C.3}$$

Equation (C.4) gives the probability of assigning the index of a subset's element to a word.

$$p(z_{d,n}|\theta_d) = \theta_{d,z_{d,n}} = \prod_{p=1}^{T}\theta_{d,p}^{I(z_{d,n}=p)} \tag{C.4}$$

Equation (C.5) gives the probability of a subset assignment to a document.

$$p(t_d|\pi) = \pi_{t_d} = \prod_{x=1}^{S}\pi_x^{I(t_d=x)} \tag{C.5}$$

Equation (C.6) gives the probability of a word given the index of subset's element and the subset associated with the document.

$$p(w_{d,n}|z_{d,n},\beta,t_d) = \beta_{t_d.z_{d,n},w_{d,n}} = \prod_{x,p,i,j}^{S,T,K,V}\beta_{i,j}^{I(w_{d,n}=j)I(z_{d,n}=p)I(t_d=x)I(x.p=i)} \tag{C.6}$$

Equations (C.7), (C.9), (C.11), (C.13) and (C.15) define the variational distributions used in the posterior approximation. Equations (C.8), (C.10), (C.12), (C.14) and (C.16) give some necessary expectations for the proof. The expectation terms from (C.8), (C.10) and (C.12) are obtained based on the observation that the first derivative of the log normalizer is equal to the expected value of the sufficient statistics, while the ones from (C.14) and (C.16) result from the fact that the expected value of the indicator of a variable taking on a particular setting is the probability of the variable being in that setting.

Equation (C.7) represents the variational distribution of the subset proportions in the exponential family form.

$$q(\pi|\mu) = Dir(\mu) = \exp\{(\sum_{x=1}^{S}(\mu_x - 1)\log\pi_x) + \log\Gamma(\mu_0) - \sum_{x=1}^{S}\log\Gamma(\mu_x)\} \tag{C.7}$$

$$E_q[\log\pi_x|\mu] = \Psi(\mu_x) - \Psi(\mu_0) \tag{C.8}$$

Equation (C.9) represents the variational distribution of a topic in the exponential family form.

$$q(\beta_i|\lambda_i) = Dir(\lambda_i) = \exp\{(\sum_{j=1}^{V}(\lambda_{i,j}-1)\log\beta_{i,j}) + \log\Gamma(\lambda_{i,0}) - \sum_{j=1}^{V}\log\Gamma(\lambda_{i,j})\} \quad \text{(C.9)}$$

$$E_q[\log\beta_{i,j}|\lambda i] = \Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}) \quad \text{(C.10)}$$

Equation (C.11) represents the variational distribution of the proportions over the indexes of a subset's elements in the exponential family form.

$$q(\theta_d|\gamma_d) = Dir(\gamma_d) = \exp\{(\sum_{p=1}^{T}(\gamma_{d,p}-1)\log\theta_{d,p}) + \log\Gamma(\gamma_{d,0}) - \sum_{p=1}^{T}\log\Gamma(\gamma_{d,p})\} \quad \text{(C.11)}$$

$$E_q[\log\theta_{d,p}|\gamma_d] = \Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0}) \quad \text{(C.12)}$$

Equation (C.13) gives the variational distribution of the assignment of an index of a subset's element to a word.

$$q(z_{d,n}|\phi_{d,n}) = Cat(\phi_{d,n}) = \phi_{d,n,z_{d,n}} = \prod_{p=1}^{T}\phi_{d,n,p}^{I(z_{d,n}=p)} \quad \text{(C.13)}$$

$$E_q[I(z_{d,n}=i)] = q(z_{d,n}=i|\phi_{d,n}) = \phi_{d,n,i} \quad \text{(C.14)}$$

Equation (C.15) gives the variational distribution of a subset assignment to a document.

$$q(t_d|\zeta_d) = Cat(\zeta_d) = \zeta_{d,t_d} = \prod_{x=1}^{S}\zeta_{d,x}^{I(t_d=x)} \quad \text{(C.15)}$$

$$E_q[I(t_d=x)] = q(t_d=x|\zeta_d) = \zeta_{d,x} \quad \text{(C.16)}$$

C.1 Deriving the Evidence Lower Bound

The ELBO is the objective function which needs to be maximized. The maximization is done using a coordinate ascent algorithm in which the variational parameters are updated iteratively until the ELBO convergences. Monitoring the value of the ELBO is useful for assessing algorithm termination, but also for sanity checks (the ELBO is guaranteed to increase with every iteration).

In Equation (C.17) we expand the lower bound according to the conditional dependencies of the model and those of the variational distribution.

$$\begin{aligned}\mathcal{L} &= E_q[\log p(\theta, \beta, \pi, w, z, t|\alpha, \eta, \delta)] - E_q[\log q(\theta, \beta, \pi, z, t|\gamma, \lambda, \mu, \phi, \zeta)] \\ &= E_q[\log p(\pi|\delta)] + E_q[\log p(\theta|\alpha)] + E_q[\log p(\beta|\eta)] + E_q[\log p(z|\theta)] + \\ &+ E_q[\log p(t|\pi)] + E_q[\log p(w|z, t, \beta)] - E_q[\log q(\theta|\gamma)] - \\ &- E_q[\log q(\beta|\lambda)] - E_q[\log q(\pi|\mu)] - E_q[\log q(z|\phi)] - E_q[\log q(t|\zeta)]\end{aligned} \tag{C.17}$$

In Equation (C.18) we derive the expectation term that regards the probability of the proportions over the indexes of a subset's elements.

$$\begin{aligned}E_q[\log p(\theta|\alpha)] &= E_q[\log \prod_{d=1}^{M} p(\theta_d|\alpha)] \\ &= E_q[\sum_{d=1}^{M} \log p(\theta_d|\alpha)] \\ &= \sum_{d,p}^{M,T} (\alpha_p - 1)(\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0})) + \sum_{d=1}^{M} \log \Gamma(\alpha_0) - \sum_{d,p}^{M,T} \log \Gamma(\alpha_p)\end{aligned} \tag{C.18}$$

In Equation (C.19) we derive the expectation term that regards the probability

of the topics.

$$\begin{aligned} E_q[\log p(\beta|\eta)] &= E_q[\log \prod_{i=1}^{K} p(\beta_i|\eta)] \\ &= E_q[\sum_{i=1}^{K} \log p(\beta_i|\eta)] \\ &= \sum_{i,j}^{K,V} (\eta_j - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\eta_0) - \sum_{i,j}^{K,V} \log \Gamma(\eta_j) \end{aligned} \tag{C.19}$$

In Equation (C.20) we derive the expectation term that regards the probability of the proportions over the subsets.

$$E_q[\log p(\pi|\delta)] = \sum_{x=1}^{S} (\delta_x - 1)(\Psi(\mu_x) - \Psi(\mu_0)) + \log \Gamma(\delta_0) - \sum_{x=1}^{S} \log \Gamma(\delta_x) \tag{C.20}$$

In Equation (C.21) we derive the expectation term that regards the probability of assigning the index of a subset's element to every word.

$$\begin{aligned} E_q[\log p(z|\theta)] &= E_q[\log \prod_{d,n}^{M,N_d} p(z_{d,n}|\theta_d)] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \theta_{d,z_{d,n}}] \\ &= E_q[\log \prod_{d,n,p}^{M,N_d,T} \theta_{d,p}^{I(z_{d,n}=p)}] \\ &= \sum_{d,n,p}^{M,N_d,T} \phi_{d,n,p}(\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0})) \end{aligned} \tag{C.21}$$

In Equation (C.22) we derive the expectation term that regards the probability

of assigning a subset to every document.

$$\begin{aligned} E_q[\log p(t|\pi)] &= E_q[\log \prod_{d=1}^{M} p(t_d|\pi)] \\ &= E_q[\log \prod_{d=1}^{M} \pi_{t_d}] \\ &= E_q[\log \prod_{d,x}^{M,S} \pi_x^{I(t_d=x)}] \\ &= \sum_{d,x}^{M,S} \zeta_{d,x}(\Psi(\mu_x) - \Psi(\mu_0)) \end{aligned} \tag{C.22}$$

In Equation (C.18) we derive the expectation term that regards the probability of the words given the subset assigned to the document and the indication of which element of the subset to draw from.

$$\begin{aligned} E_q[\log p(w|\beta, z, t)] &= E_q[\log \prod_{d,n}^{M,N_d} p(w_{d,n}|t_d, z_{d,n}, \beta)] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \beta_{t_d.z_{d,n},w_{d,n}}] \\ &= E_q[\log \prod_{d,n,x,p,i,j}^{M,N_d,S,T,K,V} \beta_{i,j}^{I(w_{d,n}=j)I(z_{d,n}=p)I(t_d=x)I(x.p=i)}] \\ &= \sum_{d,n,x,p,i,j}^{M,N_d,S,T,K,V} \zeta_{d,x}\phi_{d,n,p}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n}=j)I(x.p=i) \end{aligned} \tag{C.23}$$

In Equation (C.24) we derive the expectation term that regards the variational distributions of the proportions over a subset's elements.

$$\begin{aligned} E_q[\log q(\theta|\gamma)] &= E_q[\log \prod_{d=1}^{M} q(\theta_d|\gamma_d)] \\ &= E_q[\sum_{d=1}^{M} \log p(\theta_d|\gamma_d)] \\ &= \sum_{d,p}^{M,T} (\gamma_{d,p} - 1)(\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0})) + \sum_{d=1}^{M} \log\Gamma(\gamma_{d,0}) - \sum_{d,p}^{M,T} \log\Gamma(\gamma_{d,p}) \end{aligned} \tag{C.24}$$

In Equation (C.25) we derive the expectation term that regards the variational distributions of the topics.

$$\begin{aligned} E_q[\log q(\beta|\lambda)] &= E_q[\log \prod_{i=1}^{K} q(\beta_i|\lambda_i)] \\ &= E_q[\sum_{i=1}^{K} \log p(\beta_i|\lambda_i)] \\ &= \sum_{i,j}^{K,V} (\lambda_{i,j} - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \sum_{i=1}^{K} \log \Gamma(\lambda_{i,0}) - \sum_{i,j}^{K,V} \log \Gamma(\lambda_{i,j}) \end{aligned} \tag{C.25}$$

In Equation (C.26) we derive the expectation term that regards the variational distribution of the proportions over the subsets.

$$E_q[\log q(\pi|\mu)] = \sum_{x=1}^{S} (\mu_x - 1)(\Psi(\mu_x) - \Psi(\mu_0)) + \log \Gamma(\mu_0) - \sum_{x=1}^{S} \log \Gamma(\mu_x) \tag{C.26}$$

In Equation (C.27) we derive the expectation term that regards the variational distributions of the assignment of subset indexes to words.

$$\begin{aligned} E_q[\log q(z|\phi)] &= E_q[\log \prod_{d,n}^{M,N_d} q(z_{d,n}|\phi_{d,n})] \\ &= E_q[\log \prod_{d,n}^{M,N_d} \phi_{d,n,z_{d,n}}] \\ &= E_q[\log \prod_{d,n,p}^{M,N_d,T} \phi_{d,n,p}^{I(z_{d,n}=p)}] \\ &= \sum_{d,n,p}^{M,N_d,T} \phi_{d,n,p} \log \phi_{d,n,p} \end{aligned} \tag{C.27}$$

In Equation (C.28) we derive the expectation term that regards the variational

distributions of the assignment of subsets to documents.

$$\begin{aligned} E_q[\log q(t|\zeta)] &= E_q[\log \prod_{d=1}^{M} q(t_d|\zeta_d)] \\ &= E_q[\log \prod_{d=1}^{M} \zeta_{d,t_d}] \\ &= E_q[\log \prod_{d,x}^{M,S} \zeta_{d,x}^{I(t_d=x)}] \\ &= \sum_{d,x}^{M,S} \zeta_{d,x} \log \zeta_{d,x} \end{aligned} \tag{C.28}$$

C.2 Deriving the Update Formulas of the Variational Parameters

Maximizing the lower bound with respect to every variational parameter leads to the update formulas from Equations (C.31), (C.34), (C.37), (C.40) and (C.43).

Equations (C.29), (C.30) and (C.31) address the maximization with respect to the variational parameter corresponding to the subset proportions.

$$\begin{aligned} \mathcal{L}_{\mu_x} = (\delta_x - 1)(\Psi(\mu_x) - \Psi(\mu_0)) + (\Psi(\mu_x) - \Psi(\mu_0)) \sum_{d=1}^{M} \zeta_{d,x} - \\ - (\mu_x - 1)(\Psi(\mu_x) - \Psi(\mu_0)) - \log \Gamma(\mu_0) + \log \Gamma(\mu_x) \end{aligned} \tag{C.29}$$

$$\begin{aligned} (\mathcal{L}_{\mu_x})' = (\delta_x - 1)(\Psi'(\mu_x) - \Psi'(\mu_0)) + (\Psi'(\mu_x) - \Psi'(\mu_0)) \sum_{d=1}^{M} \zeta_{d,x} - \\ - (\mu_x - 1)(\Psi'(\mu_x) - \Psi'(\mu_0)) - \Psi(\mu_x) + \Psi(\mu_0) - \Psi(\mu_0) + \Psi(\mu_x) \\ = (\Psi'(\mu_x) - \Psi'(\mu_0))(\delta_x + \sum_{d=1}^{M} \zeta_{d,x} - \mu_x) \end{aligned} \tag{C.30}$$

$$\mu_x = \delta_x + \sum_{d=1}^{M} \zeta_{d,x} \tag{C.31}$$

Equations (C.32), (C.33) and (C.34) address the maximization with respect to the variational parameter corresponding to the document-level proportions over the

indexes of a subset.

$$\begin{aligned}\mathcal{L}_{\gamma_{d,p}} = &(\alpha_p - 1)(\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0})) + (\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0}))\sum_{n=1}^{N_d}\phi_{d,n,p} - \\ &- (\gamma_{d,p} - 1)(\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0})) - \log\Gamma(\gamma_{d,0}) + \log\Gamma(\gamma_{d,p})\end{aligned} \tag{C.32}$$

$$\begin{aligned}(\mathcal{L}_{\gamma_{d,p}})' = &(\alpha_p - 1)(\Psi'(\gamma_{d,p}) - \Psi'(\gamma_{d,0})) + (\Psi'(\gamma_{d,p}) - \Psi'(\gamma_{d,0}))\sum_{n=1}^{N_d}\phi_{d,n,p} - \\ &- (\gamma_{d,p} - 1)(\Psi'(\gamma_{d,p}) - \Psi'(\gamma_{d,0})) - \Psi(\gamma_{d,p}) + \Psi(\gamma_{d,0}) - \Psi(\gamma_{d,0}) + \Psi(\gamma_{d,p}) \\ &= (\Psi'(\gamma_{d,p}) - \Psi'(\gamma_{d,0}))(\alpha_p + \sum_{n=1}^{N_d}\phi_{d,n,p} - \gamma_{d,p})\end{aligned} \tag{C.33}$$

$$\gamma_{d,p} = \alpha_p + \sum_{n=1}^{N_d}\phi_{d,n,p} \tag{C.34}$$

Equations (C.35), (C.36) and (C.37) address the maximization with respect to the variational parameter corresponding to a topic.

$$\begin{aligned}\mathcal{L}_{\lambda_{i,j}} = &(\eta_j - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) + \\ &+ (\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))\sum_{d,n,x,p}^{M,N_d,S,T}\zeta_{d,x}\phi_{d,n,p}I(w_{d,n} = j)I(x.p = i) - \\ &- (\lambda_{i,j} - 1)(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0})) - \log\Gamma(\lambda_{i,0}) + \log\Gamma(\lambda_{i,j})\end{aligned} \tag{C.35}$$

$$\begin{aligned}(\mathcal{L}_{\lambda_{i,j}})' = &(\eta_j - 1)(\Psi'(\lambda_{i,j}) - \Psi'(\lambda_{i,0})) + \\ &+ (\Psi'(\lambda_{i,j}) - \Psi'(\lambda_{i,0}))\sum_{d,n,x,p}^{M,N_d,S,T}\zeta_{d,x}\phi_{d,n,p}I(w_{d,n} = j)I(x.p = i) - \\ &- (\lambda_{i,j} - 1)(\Psi'(\lambda_{i,j}) - \Psi'(\lambda_{i,0})) - \Psi(\lambda_{i,j}) + \Psi(\lambda_{i,0}) - \Psi(\lambda_{i,0}) + \Psi(\lambda_{i,j}) \\ &= (\Psi'(\lambda_{i,j}) - \Psi'(\lambda_{i,0}))(\eta_j + \sum_{d,n,x,p}^{M,N_d,S,T}\zeta_{d,x}\phi_{d,n,p}I(w_{d,n} = j)I(x.p = i) - \lambda_{d,p})\end{aligned} \tag{C.36}$$

$$\lambda_{i,j} = \alpha_p + \sum_{d,n,x,p}^{M,N_d,S,T} \zeta_{d,x}\phi_{d,n,p} I(w_{d,n} = j) I(x.p = i) \tag{C.37}$$

Equations (C.38), (C.39) and (C.40) address the maximization with respect to the variational parameter corresponding to the subset-index assignment to a word. This is a constrained maximization where $\sum_{p=1}^{T} \phi_{d,n,p} = 1$. The Lagrangian is presented below.

$$\begin{aligned}\mathcal{L}_{\phi_{d,n,p}} &= \phi_{d,n,p}(\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0}))+ \\ &+ \phi_{d,n,p} \sum_{x,i,j}^{S,K,V} \zeta_{d,x}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)- \\ &- \phi_{d,n,p} \log \phi_{d,n,p} + a_{d,n} \sum_{p=1}^{T}(\phi_{d,n,p} - 1)\end{aligned} \tag{C.38}$$

$$\begin{aligned}(\mathcal{L}_{\phi_{d,n,p}})' &= \Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0})+ \\ &+ \sum_{x,i,j}^{S,K,V} \zeta_{d,x}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)- \\ &- \log \phi_{d,n,p} - 1 + a_{d,n}\end{aligned} \tag{C.39}$$

$$\phi_{d,n,p} \propto \exp\{\Psi(\gamma_{d,p}) - \Psi(\gamma_{d,0}) + \sum_{x,i,j}^{S,K,V} \zeta_{d,x}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)\} \tag{C.40}$$

Equations (C.41), (C.42) and (C.43) address the maximization with respect to the variational parameter corresponding to the subset assignment to a document. This is a constrained maximization where $\sum_{x=1}^{S} \zeta_{d,x} = 1$. The Lagrangian is presented below.

$$\begin{aligned}\mathcal{L}_{\zeta_{d,x}} &= \zeta_{d,x}(\Psi(\mu_x) - \Psi(\mu_0))+ \\ &+ \zeta_{d,x} \sum_{n,p,i,j}^{N_d,T,K,V} \phi_{d,n,p}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)- \\ &- \zeta_{d,x} \log \zeta_{d,x} + b_d \sum_{x=1}^{S}(\zeta_{d,x} - 1)\end{aligned} \tag{C.41}$$

$$
\begin{aligned}
(\mathcal{L}_{\zeta_{d,x}})' = \Psi(\mu_x) - \Psi(\mu_0) + \\
+ \sum_{n,p,i,j}^{N_d,T,K,V} \phi_{d,n,p}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i) - \\
- \log \zeta_{d,x} - 1 + b_d
\end{aligned}
\tag{C.42}
$$

$$
\zeta_{d,x} \propto \exp\{\Psi(\mu_x) - \Psi(\mu_0) + \sum_{n,p,i,j}^{N_d,T,K,V} \phi_{d,n,p}(\Psi(\lambda_{i,j}) - \Psi(\lambda_{i,0}))I(w_{d,n} = j)I(x.p = i)\}
\tag{C.43}
$$

References

[1] Abramowitz, M. (1974). *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables,*. Dover Publications, Incorporated. 85, 101, 109

[2] AlSumait, L., Barbar, D. & Domeniconi, C. (2008). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *2008 Eighth IEEE International Conference on Data Mining*, 3–12. 19

[3] Asuncion, A., Welling, M., Smyth, P. & Teh, Y.W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, 27–34, AUAI Press, Arlington, Virginia, United States. 16

[4] Blei, D. & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, **18**, 147. 2, 10, 81

[5] Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, **55**, 77–84. 1, 7, 9

[6] Blei, D.M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, **1**, 203–232. 1, 7, 9

[7] Blei, D.M. & Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 113–120, ACM, New York, NY, USA. 2, 10, 16, 81

[8] Blei, D.M. & Lafferty, J.D. (2009). Topic models. *Text mining: classification, clustering, and applications*, **10**, 34. 1, 7, 9, 34, 56, 74

[9] Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**, 993–1022. 1, 9, 12, 16, 17, 19, 22, 54

[10] Blei, D.M., Kucukelbir, A. & McAuliffe, J.D. (2016). Variational Inference: A Review for Statisticians. *ArXiv e-prints*. 16, 17, 30, 72

[11] Cardoso-Cachopo, A. (2007). Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa. 74

[12] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. & Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, 288–296. 19, 20

[13] Chen, W., Wang, J., Zhang, Y., Yan, H. & Li, X. (2015). User based aggregation for biterm topic model. *Volume 2: Short Papers*, 489. 13, 22

[14] Das, R., Zaheer, M. & Dyer, C. (2015). Gaussian lda for topic models with word embeddings. In *ACL (1)*, 795–804. 11

[15] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R. & Lin, C.J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, **9**, 1871–1874. 22

[16] Griffiths, T.L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, **101**, 5228–5235. 9

[17] Griffiths, T.L., Steyvers, M., Blei, D.M. & Tenenbaum, J.B. (2005). Integrating topics and syntax. In L.K. Saul, Y. Weiss & L. Bottou, eds., *Advances in Neural Information Processing Systems 17*, 537–544, MIT Press. 10

[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, **11**, 10–18. 22

[19] Hoffman, M.D., Blei, D.M. & Bach, F. (2010). Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, 856–864, Curran Associates Inc., USA. 16, 19, 81

[20] Hoffman, M.D., Blei, D.M., Wang, C. & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, **14**, 1303–1347. 10, 16, 17, 18, 19, 26, 50, 81

[21] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, 289–296, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 9, 12

[22] Hong, L. & Davison, B.D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, 80–88, ACM, New York, NY, USA. 2, 3, 13, 19, 20, 24, 33, 35, 47

[23] Kruschke, J.K. (2015). Chapter 7 - markov chain monte carlo. In J.K. Kruschke, ed., *Doing Bayesian Data Analysis (Second Edition)*, 143 – 191, Academic Press, Boston, second edition edn. 16

[24] Li, J., Liao, M., Gao, W., He, Y. & Wong, K.F. (2016). Topic extraction from microblog posts using conversation structures. In *ACL (1)*, The Association for Computer Linguistics. 12, 15, 20

[25] Li, W. & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 577–584, ACM, New York, NY, USA. 2, 7, 10, 19, 81

[26] Lin, T., Tian, W., Mei, Q. & Cheng, H. (2014). The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings*

of the 23rd International Conference on World Wide Web, WWW '14, 539–550, ACM, New York, NY, USA. 2, 3, 13, 14, 19, 20, 22

[27] Lu, Y., Mei, Q. & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, **14**, 178–203. 2

[28] Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. 21, 22

[29] McCallum, A., Wang, X. & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, **30**, 249–272. 12

[30] McCallum, A.K. (1999). Multi-label text classification with a mixture model trained by em. In *AAAI 99 Workshop on Text Learning*. 12

[31] Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, 889–892, ACM, New York, NY, USA. 2, 3, 13, 19, 20, 24, 33, 35, 47

[32] Mei, Q. & Zhai, C. (2006). A mixture model for contextual text mining. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, 649–655, ACM, New York, NY, USA. 7, 11, 12, 81

[33] Mei, Q., Liu, C., Su, H. & Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, 533–542, ACM, New York, NY, USA. 2, 11, 12, 81

[34] Mimno, D., Wallach, H.M., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the*

Conference on Empirical Methods in Natural Language Processing, EMNLP '11, 262–272, Association for Computational Linguistics, Stroudsburg, PA, USA. 20

[35] Newman, D., Lau, J.H., Grieser, K. & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 100–108, Association for Computational Linguistics, Stroudsburg, PA, USA. 19, 20

[36] Nguyen, D.Q., Billingsley, R., Du, L. & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, **3**, 299–313. 3, 8, 11, 12, 13, 16, 20, 24, 33, 56, 66, 73

[37] Nigam, K., Mccallum, A.K., Thrun, S. & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, **39**, 103–134. 8

[38] Petrović, S., Osborne, M. & Lavrenko, V. (2012). Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, 338–346, Association for Computational Linguistics, Stroudsburg, PA, USA. 57

[39] Quan, X., Kit, C., Ge, Y. & Pan, S.J. (2015). Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 2270–2276, AAAI Press. 2, 14, 15, 22

[40] Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, 487–494, AUAI Press, Arlington, Virginia, United States. 2, 12, 81

[41] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, **28**, 4:1–4:38. 7, 12

[42] Sasaki, K., Yoshikawa, T. & Furuhashi, T. (2014). Online topic model for twitter considering dynamics of user interests and topic trends. In *EMNLP*, 1977–1985. 13

[43] Sato, I. & Nakagawa, H. (2012). Rethinking Collapsed Variational Bayes Inference for LDA. *ArXiv e-prints*. 16

[44] Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, **427**, 424–440. 9

[45] Tang, J., Meng, Z., Nguyen, X., Mei, Q. & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *ICML*, 190–198. 12, 66

[46] Teh, Y.W., Newman, D. & Welling, M. (2006). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, vol. 6, 1378–1385. 16

[47] Teh, Y.W., Kurihara, K. & Welling, M. (2007). Collapsed variational inference for hdp. In *Advances in neural information processing systems*, 1481–1488. 16

[48] Vitale, D., Ferragina, P. & Scaiella, U. (2012). Classification of short texts by deploying topical annotations. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, 376–387, Springer-Verlag, Berlin, Heidelberg. 74

[49] Vosecky, J., Jiang, D., Leung, K.W.T. & Ng, W. (2013). Dynamic multi-faceted topic discovery in twitter. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, 879–884, ACM, New York, NY, USA. 2, 13, 15, 20, 24

[50] Vosecky, J., Jiang, D., Leung, K.W.T., Xing, K. & Ng, W. (2014). Integrating social and auxiliary semantics for multifaceted topic modeling in twitter. *ACM Transactions on Internet Technology*, **14**, 27:1–27:24. 2, 13, 15, 20, 24, 25

[51] Wallach, H.M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 977–984, ACM, New York, NY, USA. 10

[52] Wallach, H.M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 1105–1112, ACM, New York, NY, USA. 19

[53] Wang, C., Paisley, J.W. & Blei, D.M. (2011). Online variational inference for the hierarchical dirichlet process. In *AISTATS*, vol. 2, 4. 10

[54] Wang, X. & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, 424–433, ACM, New York, NY, USA. 10, 81

[55] Weng, J., Lim, E.P., Jiang, J. & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, 261–270, ACM, New York, NY, USA. 13, 24

[56] Yan, X., Guo, J., Lan, Y. & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 1445–1456, ACM, New York, NY, USA. 2, 3, 4, 8, 12, 13, 15, 16, 19, 20, 22, 24, 33, 34, 46, 50, 56, 66, 73, 74

[57] Yan, X., Guo, J., Lan, Y., Xu, J. & Cheng, X. (2015). A probabilistic model for bursty topic discovery in microblogs. In *Proceedings of the Twenty-*

Ninth AAAI Conference on Artificial Intelligence, AAAI'15, 353–359, AAAI Press. 13

[58] Zhai, C., Velivelli, A. & Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 743–748, ACM, New York, NY, USA. 11, 12

[59] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H. & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, 338–349, Springer-Verlag, Berlin, Heidelberg. 3, 13, 25

[60] Zhou, D., Chen, L. & He, Y. (2014). A simple bayesian modelling approach to event extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 700–705, Association for Computational Linguistics, Baltimore, Maryland. 14

[61] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K. & Xiong, H. (2016). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 2105–2114, ACM, New York, NY, USA. 2, 3, 12, 13, 14