Online ET-LDA - Joint Modeling of Events and their Related Tweets with Online Streaming Data

by

Anirudh Acharya

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2015 by the
Graduate Supervisory Committee:

Subbarao Kambhampati, Chair
Hasan Davulcu
Hanghang Tong

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Micro-blogging platforms like Twitter have become some of the most popular sites for people to share and express their views and opinions about public events like debates, sports events or other news articles. These social updates by people complement the written news articles or transcripts of events in giving the popular public opinion about these events. So it would be useful to annotate the transcript with tweets. The technical challenge is to align the tweets with the correct segment of the transcript. ET-LDA by Hu *et al* [9] addresses this issue by modeling the whole process with an LDA-based graphical model. The system segments the transcript into coherent and meaningful parts and also determines if a tweet is a general tweet about the event or it refers to a particular segment of the transcript. One characteristic of the Hu *et al*'s model is that it expects all the data to be available upfront and uses batch inference procedure. But in many cases we find that data is not available beforehand, and it is often streaming. In such cases it is infeasible to repeatedly run the batch inference algorithm. My thesis presents an online inference algorithm for the ET-LDA model, with a continuous stream of tweet data and compare their runtime and performance to existing algorithms.

ACKNOWLEDGEMENTS

I would foremost like to thank my advisor Prof. Subbarao Kambhampati for giving me the opportunity to work in his lab. I would like to thank him for his guidance and immense patience with which he advised me.

I would also like to express my gratitude to my committee members Dr. Hasan Davulcu and Dr Hanghang Tong for taking time out and agreeing to be a part of my defense committee.

I would like to thank Dr Yuheng Hu for helping me and guiding me throughout my research. Discussions with him helped a great deal in not only refining my idea but also in evaluating them.

I would like to thank Mrs Farideh Tadayon-Navabi, and Dr Jingrui He for supporting my education with Teaching Assistantships.

I am also extremely grateful to all my friends, lab-mates and other staff members who supported and helped me during my stay in Arizona State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Social Media and Microblogging platforms are becoming an integral part of journalism and news consumption by the public. Twitter, a microblogging platform, as of today, has over 302 million monthly active users[1]. Events in countries like Egypt and Tunisia, where huge changes in the political landscape of the countries was facilitated by twitter as the main tool of communication, demonstrates the effectiveness of social media as a platform to disseminate information and in expressing and galvanizing opinions about events. Such microblogging platforms are also fast becoming the venue for people to discuss their opinions and share their thoughts about public events such as Presidential debates and speeches. Indeed, often there is a spike in the usage of these platforms when such events are occurring. These events also show the effectiveness of social media in covering news events that are not covered in conventional media. In fact it is not uncommon these days for conventional media to pick up news stories from the trending topics of the day from social media.

A popular trend emerging in journalistic circles is to combine news stories from conventional media with posts and updates from social media to enrich the experience of news consumption. Annotating news stories with social media updates ensures that the readers get an idea as to how that particular bit of news was received by the general public or how certain prominent members of the society reacted to that particular bit of news. With the huge amount of data generated by social media, annotating news stories with social media posts cannot always be done manually and automated solutions will be required. It is also common practice for many political organizations to run social media campaigns about the causes they care about. Automated solutions would also be useful here.

Twitter is one such platform where people share news stories and also voice their opinion on various matters. The usage of the platform peaks during public events like debates and speeches, when people use it to not only voice their own opinions, but also to view the opinion of other people and communicate with them. This platform acts as a virtual town square where people discuss various events. Engaging in this platform enriches the experience of the events like the Presidential Debate or Speeches. This also serves as a source of data, like a survey or a poll, to measure public sentiment about the events. This creates an opportunity for journalists and other political scientists to gauge the public mood about different aspects discussed in the debate.

The technical challenge here is to retrieve tweets relevant to particular segment of the story, in other words to align tweets with the respective parts of the story which they refer to in a meaningful way. ET-LDA by Hu *et al* [9] attempted this problem by modeling the problem as a LDA-based probabilistic graphical model with a generative story, and then infer the parameters using approximate inference techniques like Markov Chain Monte Carlo.

ET-LDA is a joint topic model to segment the transcript and also infer the topic distributions of the different segments and the segment indicators of the tweets. ET-LDA is different from earlier works in that, it takes the topical influence of tweets into account while jointly modeling the transcript segmentation and the topic-word distributions.

ET-LDA however requires the complete dataset to be present to run the inference algorithm, in other terms it is a batch system. But Twitter is a live stream of data and in many situations like a live event, we do not have a complete corpus of tweets for us to run the ET-LDA's batch inference. Hence the need for an online inference algorithm, which infers and updates the model parameters as we obtain more data from the twitter stream. Although online inference algorithms exist for graphical models like Latent Dirichlet Allocation [6], they need to be adapted and modified to work for the joint model of ET-LDA.

The contribution of the thesis is developing an online inference procedure for the ET-LDA graphical model. This involved altering the update rules for the parameter estimation of batch inference procedure, and building a system that is able to handle the online stream of data and continuously update the values of the latent variables. We also provide a systematic comparison of online and batch versions of ET-LDA. In particular we made a comparison of runtime of the online inference algorithm to the batch algorithm. We also compared the accuracy of the parameter estimates to those obtained by the batch version and how well they converge to the true values of the latent variables.

This thesis is organized in the following manner, in chapter 2 we describe the literature related to ET-LDA, such as text segmentation, event analysis, and annotating news articles with related social media updates. In chapter 3 we give a brief overview of the ET-LDA system. In chapter 4 we describe the online inference algorithm and its implementation. In chapter 5 we describe the experiments and the evaluation of our system. In chapter 6 we give the concluding remarks and possible extensions of the work.

Chapter 2

RELATED WORK

Online ET-LDA is an extension of the existing work on ET-LDA [9]. ET-LDA deals with finding the dominating topics of the different segments of the transcript and the individual tweets. Simultaneously it also segments the transcript and aligns the tweets with a segment of the transcript based on the topic distributions of the tweet and the segment. The graphical model of ET-LDA is an extension of the Latent Dirichlet Allocation and jointly models, finding latent topics and finding segment boundaries in the transcript. This work is an extension of topic modeling techniques like Latent Dirichlet Allocation by Blei em et al [5], which finds latent topics in a given set of documents. Blei's work uses Variational Bayes for the inference procedure, which gives an exact analytical solution to an approximation of the posterior. Griffith em et al [7] proposed a Gibbs sampling procedure for the same model, which is a Monte Carlo technique for providing a numerical approximation to the exact posterior distribution through sampling.

One of the other appealing aspects of our work is the thematic alignment of tweets and segments. Shraer *et al* [12] propose a method of annotating news articles with social media updates using a publish-subscribe pattern. Barnard *et al* [3] propose MOM-LDA which is a joint graphical model to align image segments with their relevant text annotations.

A related direction of research, one which has influenced our work to a considerable extent, has been to extend LDA to a continuous stream of online data. Online LDA has been approached with different methods. Hoffman *et al* [8] proposed an online Variational Bayes inference algorithm, which they show converges to a local optimum of the Variation Bayes objective function. Canini *et al* [6] proposed an online procedure for the Gibbs sampler, an iterative process to update the model parameters with a periodic resampling

of a subset of the old data. It had results comparable to the batch Gibbs sampler. There has also been research to extend Latent Dirichlet Allocation to short texts such as tweets. Sahami and Heilman in [11] propose a method of augmenting the tweets which has been used in this thesis.

ET-LDA by Hu *et al* [9] jointly models the problems of text segmentation and finding topic distributions to find the segment boundaries, topic distribution and the tweet transcript alignment, but does it with offline data. But an expectation of the model is to have the whole transcript and the twitter corpus beforehand to run the inference algorithm. In this thesis we propose an algorithm to do the inference with an online stream of twitter data.

Chapter 3

ET-LDA OVERVIEW

This chapter gives a brief overview of ET-LDA [9]. ET-LDA is a joint topic model to segment the transcript and align the tweets with transcript segments based on the similarity of their topics.

The graphical model of the ET-LDA is shown in Figure 3.1 and the symbol description are shown in Figure 3.2

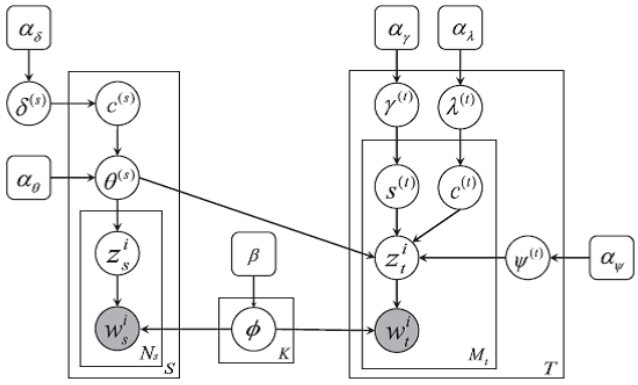

Figure 3.1: Graphical Model for ET-LDA

The generative story of the ET-LDA has been outlined in Algorithm 2. Conceptually ET-LDA tries to jointly model the event and the tweet, with the assumption that the event influences the topics of the tweets. It is also assumed that a given tweet can be associated with any segment of the event transcript, and is not restricted by any time window. Also tweets could be a specific tweet, meaning the subject matter of the tweet is pertaining to one particular segment of the transcript, or it could be a general tweet, meaning the content of the tweet is about the whole event in general. ET-LDA also segments splits the transcript

Notation	Description
S	a set of paragraphs in the event's transcript
N_s	the number of words in paragraph s
T	a set of tweets associated with the event
M_t	the number of words in tweet t
$\theta^{(s)}$	topic mixture of the specific topics from a paragraph s of the event
$\psi^{(t)}$	topic mixture of the general topics from tweets corpus
$\delta^{(s)}$	parameter for choosing to draw topics in paragraph s from $\theta^{(s)}$ or $\theta^{(s-1)}$
$c^{(s)}$	indicates whether the topic of a paragraph is drawn from current or previous segment's topics.
$\lambda^{(t)}$	parameter for choosing to draw topics in t from θ or ψ
$c^{(t)}$	indicates whether the topic of a tweet is drawn from specific or general topics
$s^{(t)}$	a referred segment, to which a specific topic in a tweet is associated
w_s, w_t	words in event's transcript, tweets, respectively
z_s, z_t	topic assignments of words in event, tweets, respectively.
α, β	Dirichlet/beta parameters of the Multinomial/Bernoulli distributions

Figure 3.2: Symbol Description for ET-LDA Graphical Model

into sequential segments with each segment covering a particular set of topics. The latent variables of the model inferred using approximate techniques like Markov Chain Monte Carlo.

3.1 Gibbs Sampling

Gibbs sampling also called Markov Chain Monte Carlo is used when the joint distribution is not known explicitly or it is hard to sample from it directly, but the conditional distribution of each variable is known and we can sample from these conditional distributions[2]. Gibbs sampling is only applicable when the random variable X has atleast two dimensions, i.e. each point x is actually $x =< x_1, x_2, ..., x_3 >$ with $k > 1$.

The basic idea behind Gibbs sampling is that instead of probabilistically picking the

Algorithm 1 Generative Model of ET-LDA

for every para $s \in S$ **do**
  draw a segment choice indicator $c^{(s)} \sim Bernoulli(\delta^{(s)})$
  if $c^{(s)} = 1$ **then**
    draw a topic mixture $\theta^{(s)} \sim Dirichlet(\alpha_\theta)$
  else
    draw a topic mixture $\theta^{(s)} \sim \delta(\theta^{(s-1)}, \theta^{(s)})$
  end if
  for each word $w_s^i \in s$ **do**
    draw a topic $z_s^i \sim Multinomial(\theta^{(s)})$
    draw a word $w_s^i \sim \phi_{z_s^i}$
  end for
end for
for every tweet $t \in T$ **do**
  for each word $w_t^i \in t$ **do**
    draw a topic changing indicator $c^{(t)} \sim Bernoulli(\lambda^{(t)})$
    if $c^(t) = 1$ **then**
      draw a topic mixture $\psi^{(t)} \sim Dirichlet(\alpha_\psi)$
      draw a general topic $z_t^i \sim Multinomial(\theta^{(s)})$
    else
      draw a para indicator $s \sim Categorical(\gamma^{(t)})$
      draw a specific topic $z_t^i \sim Multinomial(\theta^{(s)})$
    end if
  end for
end for

value of the variable at the next time iteration $x_{(t+1)}$ directly, we sample each of the k dimensions separately, conditioned on the other $k-1$ dimensions. Such a sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is the desired joint distribution [2]. The generic Gibbs Sampling algorithm is as shown in Algorithm 2.

Algorithm 2 Gibbs Sampling Algorithm

1: $x^{(0)} =< x_1^{(0)}, x_2^{(0)}, ...x_k^{(0)} >$
2: **for** t=1 to T **do**
3: **for** i=1 to k **do**
4: $x_i^{(t+1)} \sim P(X_i|x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, ..., x_k^{(t)})$
5: **end for**
6: **end for**

Sampling the posterior distribution of a Bayesian network is specifically well suited to Gibbs Sampling, since Bayesian networks are typically specified as a collection of conditional distributions.

3.2 Collapsed Gibbs Sampling

A collapsed Gibbs sampler means that we marginalize the joint distribution over one or more variables and then we sample from the distribution. For example if we have a distribution $P(x, y, z)$ we can computer $P(x, y)$ by summing over all possible values of z. Now though we have accounted for the variable z we do not have to deal with the complexities of manipulating it explicitly as a parameter for every iteration of the sampling process. If z, instead of a discrete variable is a continuous variable then we will be integrating over all values of z instead of summing. Also if we are using conjugate priors, with distributions belonging to the same exponential family, which we do both in LDA and ET-LDA,

then the inference rules for the Gibbs sampler will become considerably simpler by using a collapsed Gibbs sampling, where we integrate out the conjugate prior, as opposed to just Gibbs sampling[10].

3.3 Batch Inference

The batch inference is shown in Algorithm 3. In batch inference we require the Gibbs sampler to read the entire data into the memory and iteratively run the Gibbs sampler over the entire dataset for every iteration. The symbol descriptions for the algorithm is given in Figure 3.2

Algorithm 3 Online Gibbs Sampling Algorithm for ET-LDA

Randomly assign the topic indicators $z_{s,N}$ *and* $z_{t,N} \sim \{1, 2, 3, ..., k\}$ *and the segment indicators* c_t, c_S, s_t

for t = 1 to T **do**

  for every para p in transcript S **do**

    // Sample the segment indicator for every para p to decide if the para will begin a new segment or it will continue in the previous segment

    $c_s \sim P(c_s|c_{-(s,i)}, c_t, w_s, w_t, z_s, z_t, s_t)$

    // Sample topic indicator for every word i in the current paragraph

    $z_{s,i} \sim P(z_{s,i}|z_{-(s,i)}, z_t, w_s, w_t, c_s, c_t, s_t)$

    // Update the distributions $\theta^{(s)}$, $c^{(s)}$, ϕ *based on the new topic assignments*

  end for

  for word i in all tweets **do**

    // Sample the general/specific distribution indicator

    $c_t \sim P(c_{(t,i)}|c_{-(t,i)}, c_s, w_s, w_t, z_s, z_t, s_t)$

    // Sample the segment indicator for the specific distribution conditioned on c_t

    $s_t \sim P(s_t|s_{-(t,i)}, c_t, c_s, w_t, w_s, z_t, z_s)$

    // Sample the topic indicator for the word

    $z_t \sim P(z_{t,i}|z_{-(t,i)}, z_s, w_s, w_t, c_s, c_t, s_t)$

    // Update the distributions $\psi^{(t)}$, $\lambda^{(t)}$, $\gamma^{(t)}$, ϕ *based on the new topic assignments*

  end for

end for

Chapter 4

ONLINE INFERENCE

In the previous chapter we gave a basic overview of the ET-LDA for learning the segmentation of the transcript, the topic distributions of the segments and tweets, and the transcript-tweet alignment. But the batch inference procedure for ET-LDA does not provide us with a way of inferring the above latent variables with online data. It does not provide us a way of updating our latent variables in the light of new data.

In this chapter we outline the online inference algorithm that we developed to handle a continuous stream of data.

4.1 Sampling Algorithm

The online inference algorithm updates the latent variables of the graphical model in the light of new data without going through all the past data. The assumption with the proposed algorithm is that the event transcript is available beforehand but the twitter stream arrives online. We obtain new tweets with each iteration of the time.

The basic idea of the algorithm is, at each iteration we run the Gibbs sampler assuming the transcript and the tweets that have arrived at that particular iteration as the whole dataset. But as this forms a very small subset of the entire dataset, our estimate of the latent distributions are bound to be off by a margin that is not insignificant. To offset this we periodically reassign topic and segment indicators for a subset of the past tweets. If this reassignment of topic and segment indicators are done often enough then the latent variables being inferred will converge to the true distribution [6].

In fact re-assigning the topics to past data too often and to the whole dataset rather than a subset of the dataset will be like running the batch algorithm over and over again. But

this would not be optimal performance, hence our online inference algorithm avoids this pitfall at the same time is able to converge to the true latent distribution values. The Online inference procedure is outlined in Algorithm 1

The following table gives a description of the different symbols used in the update rules.

n_{sw}^{k}	Number of time topic k is assigned to word w in event transcript
n_{tw}^{k}	Number of times topic k is assigned to word w in tweets
$n_{k}^{S_i}$	Number of times topic k appears in segment S_i
$nt_{k}^{S_i}$	Number of times topic k appears in tweets that align with segment S_i
n_{k}^{i}	Number of times topic k appears in general distribution
M_{t}^{0}	Number of words that belong to a specific distribution in tweet t
M_{t}^{1}	Number of words that belong to a general distribution in tweet t
n_{s}^{i}	Number of times segment s is referred to in tweet t
S_{s}^{1}	Number of times the paragraph topic changes, i.e. number of times c_s = 1
S_{s}^{0}	Number of times the paragraph topic changes, i.e. number of times c_s = 0

We will now elaborate on each of the update rules and how we will sample from each of the probability distribution.

The following rule determines the binomial distribution from which we will sample to determine if we are to start a new segment in the event transcript or will we continue with the same distribution as previous paragraph. The distribution is sampled for every paragraph of the transcript. If c_s is 0 then it will merge into the same segment as the last paragraph and if it is 1 then it will branch off to begin a new segment with its own topic distribution.

Algorithm 4 Online Gibbs Sampling Algorithm for ET-LDA

Randomly assign the topic indicators $z_{s,N}$ *and* $z_{t,N} \sim \{1, 2, 3, ..., k\}$ *and the segment indicators* c_t, c_S, s_t

for t = 1 to T **do**

  for every para p in transcript S **do**

    // Sample the segment indicator for every para p to decide if the para will begin a new segment or it will continue in the previous segment

    $c_s \sim P(c_s|c_{-(s,i)}, c_t, w_s, w_t, z_s, z_t, s_t)$

    // Sample topic indicator for every word i in the current paragraph

    $z_{s,i} \sim P(z_{s,i}|z_{-(s,i)}, z_t, w_s, w_t, c_s, c_t, s_t)$

    // Update the relevant distributions based on the new topic assignments

  end for

  Mark the current tweet for Re-sampling with probability(Rejuvenation) 0.1

  for word i in tweets of the current time epoch **do**

    // Sample the general/specific distribution indicator

    $c_t \sim P(c_{(t,i)}|c_{-(t,i)}, c_s, w_s, w_t, z_s, z_t, s_t)$

    // Sample the segment indicator for the specific distribution conditioned on c_t

    $s_t \sim P(s_t|s_{-(t,i)}, c_t, c_s, w_t, w_s, z_t, z_s)$

    // Sample the topic indicator for the word

    $z_t \sim P(z_{t,i}|z_{-(t,i)}, z_s, w_s, w_t, c_s, c_t, s_t)$

    // Update the relevant distributions based on the new topic assignments

  end for

  if *number of marked tweets crosses threshold* **then**

    Run the Gibbs sampler for the marked tweets

    Clear all marked tweets

  end if

end for

$$P(c_s|c_{-(s,i)}, c_t, w_s, w_t, z_s, z_t, s_t) \propto$$
$$\begin{cases} \frac{S_s^0+\alpha_\delta-1}{S+2\alpha_\delta-1} \times \frac{n_k^{S_i}+nt_k^{S_i}+\alpha_\theta}{n_{(.)}^{S_i}+nt_{(.)}^{S_i}+K\alpha_\theta} & \text{if } c_s = 0 \\ \frac{S_s^1+\alpha_(\delta)-1}{S+2\alpha_\delta-1} \times \frac{\Gamma(K\alpha_\theta)}{\Gamma(\theta)^K} \times \frac{\prod_{k=1}^{K}\Gamma(n_k^{S(s-1)}+nt_k^{S(s-1)}+\alpha_\theta)}{\Gamma(n_{(.)}^{S(s-1)}+nt_{(.)}^{S(s-1)}+K\alpha_\theta)} \times \frac{\prod_{k=1}^{K}\Gamma(n_k^{S(s)}+nt_k^{S(s)}+\alpha_\theta)}{\Gamma(n_{(.)}^{S(s)}+nt_{(.)}^{S(s)}+K\alpha_\theta)} & \text{if } c_s = 1 \end{cases} \tag{4.1}$$

$$P(z_{s,i}|z_{-(s,i)}, z_t, w_s, w_t, c_s, c_t, s_t) = \frac{n_{sw}^k + n_{tw}^k + \beta - 1}{n_{s(.)}^k + n_{t(.)}^k + W\beta - 1} \times \frac{n_k^{S_i} + nt_k^{S_i} + \alpha_\theta - 1}{n_{(.)}^{S_i} + nt_{(.)}^{S_i} + K\alpha_\theta - 1} \tag{4.2}$$

Intuitively, formula 2 can be seen as product of two probabilities, the left term represents the probability of word w having topic k, the right component is about the probability of topic k appearing in that segment. The product of the two will give us the probability of topic k being assigned to word w which has occurred in segment s.

$$P(z_{t,i}|z_{-(t,i)}, z_s, w_s, w_t, c_s, c_t, s_t) = \begin{cases} \frac{n_{sw}^k+n_{tw}^k+\beta-1}{n_{s(.)}^k+n_{t(.)}^k+W\beta-1} \times \frac{n_k^{S_i}+nt_k^{S_i}+\alpha_\theta-1}{n_{(.)}^{S_i}+nt_{(.)}^{S_i}+K\alpha_\theta-1} & \text{if } c_t = 0 \\ \frac{n_{sw}^k+n_{tw}^k+\beta-1}{n_{s(.)}^k+n_{t(.)}^k+W\beta-1} \times \frac{n_k^i+\alpha_\psi-1}{n_{(.)+K\alpha_\psi-1}^i} & \text{if } c_t = 1 \end{cases} \tag{4.3}$$

The above update rule for assigning the topics to the different words of the tweet can be seen as follows - If tweet is a general tweet i.e. $c_t = 1$, the probability distribution for the topic indicator is a product of two expressions, the probability of a particular word w being assigned the topic k and the probability of topic k being in the general distribution. If the tweet were a specific tweet pointing to a segment s then the update rule is same as that of assigning a topic to a word in the segment s of the event transcript. This is not surprising as the topic distributions for the tweet and the segment would be the same if the tweet is aligning with that particular segment of the transcript.

We should note that the word counts in the matrix $nt_k^{S_i}$, which maintains a topic-wise and segment-wise word count of the twitter data, continuously increases as we see more data. Unlike the batch algorithm words are not just being reassigned to different topics but rather new words that are streaming in are being assigned topics and hence progressively increasing the word counts in our matrices.

Another, rather more interesting, point to note here is how Bayesian learning fits very well into online learning. The expressions that determine the probabilities of topic assignment to words in tweets, and the expression that determines the segment-tweet alignment are at first very strongly determined by the prior values that we assign, and as it progressively learns from data, we observe from the update rules that the posterior we obtain from a particular time step will act as the prior for the next time step.

$$P(c_{(t,i)}|c_{-(t,i)}, c_s, w_s, w_t, z_s, z_t, s_t) = \begin{cases} \frac{M_t^0+\alpha_(\lambda_\lambda)-1}{M_t+\alpha_(\lambda_\gamma)+\alpha_(\lambda_\psi)-1} \times \frac{n_k^{S_i}+nt_k^{S_i}+\alpha_\theta-1}{n_{(.)}^{S_i}+nt_{(.)}^{S_i}+K\alpha_\theta-1} & \text{if } c_t = 0 \\ \frac{M_t^1+\alpha_(\lambda_\lambda-1)}{M_t+\alpha_(\lambda_\gamma)+\alpha_(\lambda_\psi)-1} \times \frac{n_k^i+\alpha_\psi-1}{n_{(.)+K\alpha_\psi-1}^i} & \text{if } c_t = 1 \end{cases} \tag{4.4}$$

The above rule is to determine of the new tweet that has arrived is a general or a specific tweet. Conditioned on the result of sampling from the above binomial distribution we either sample a topic from the general distribution or determine an aligning segment from the event transcript from the rule given below and sample a topic from the topic distribution of the corresponding segment.

$$P(s_t|s_{-(t,i)}, c_t, c_s, w_t, w_s, z_t, z_s) = \frac{n_k^{S_i} + nt_k^{S_i} + \alpha_\theta - 1}{n_{(.)}^{S_i} + nt_{(.)}^{S_i} + K\alpha_\theta - 1} \times \frac{n_s^i + \alpha_\gamma - 1}{n_{(.)}^i + S\alpha_\gamma - 1} \tag{4.5}$$

4.2 Implementation of the Inference Engine

The inference engine maintains two $K \times N$ matrix(topics by words), one for the event transcript the other for the tweets. It also maintains a $P \times K$ matrix to keep track of the

para-topic word count and a P dimensional vector for maintaining the segment boundaries in the event. The two together can determine the various segment-topic distributions. The engine also keeps track of a $T \times 2$ matrix to keep count of general and specific tweets. Also a 50 dimensional array that keeps track of the tweets that have been marked for re-sampling/rejuvenation is maintained. With these matrices and counts, we are able to determine all the latent distributions, $\theta^{(s)}$, $c^{(s)}$, $s^{(t)}$, $c^{(t)}$, $\psi^{(t)}$ and ϕ.

Chapter 5

EXPERIMENTS AND EVALUATION WITH REAL AND SYNTHETIC DATA

In this chapter we outline the experiments we ran and the evaluations we made. We will demonstrate that the latent variables of the graphical model such as the topics distributions and the segment indicators can be learnt efficiently from the online algorithm introduced in the previous chapter.

5.1 Datasets

We used two datasets for our experiments, the Denver Debate dataset, mined from the twitter API, and a synthetic corpus that we generated, which consists of a synthetically generated event transcript and the tweets associated with it. The Denver debate data consists of tweets pertaining to the debate crawled from the twitter API, and the even transcript of the debate[1]. The transcript contains 360 paragraphs. A segment in a transcript can be one or more contiguous paragraphs. Both the transcript and the tweets were preprocessed to strip the text of all the stopwords, non-English words, punctuations and other miscellaneous words like retweets and hyperlinks to urls etc. Tweets are typically texts with very few words, and once we remove the stopwords and stem the existing text, we end up with even fewer words, and as shown by Tang et al in [13] topic modeling techniques like LDA do not work very well with short texts. To address this issue we augment the tweets with additional text by the method outlined by Sahami et al in [11].

The purpose of synthetic data is that, often there could be a dissonance between the generative model we have assumed and the real event. This could cascade and affect the

[1]The event transcript was obtained from NYTimes at the following link, http://www.nytimes.com/2012/10/03/us/politics/transcript-of-the-first-presidential-debate-in-denver.html

accuracy of the inference algorithms that we are trying to evaluate. With synthetic data we will not have such issues and we can make a completely objective measure of the performance of the algorithm and we also get to observe the convergence of distributions more clearly and accurately.

The synthetic data was created using the generative model of ET-LDA. We took a dictionary of 5000 words and defined 10 multinomials over these words, which are the 10 topics of our data. A mixture of these topics defines a segment-topic distribution or a general distribution. These distributions were used to generate the event transcript and the tweet corpus.

The size of the generated transcript was 320 paragraphs which were divided into 10 segments and 10,000 tweets of 100 words each.

5.2 Tuning the Hyper-parameters

In Bayesian statistics a hyperparameter is a parameter of the prior distribution. Tuning a hyperparameter usually involves choosing a hyperparameter that effectively captures the prior knowledge we have of the system and also improve the performance of the learning or inference algorithm. So the parameter of the prior dirichlet and beta distributions of the ET-LDA model become the hyperparameters of the model.

The hyperparameters of the ET-LDA model are α_δ and α_λ for the beta priors $\delta^{(s)}$ and $\lambda^{(t)}$, α_θ, α_ψ and β for the dirichlet priors $\theta^{(s)}$, $\psi^{(t)}$and ϕ.

All the above hyperparameters are fractional parameters to symmetric priors with the exception of α_λ which is an asymmetric prior to decide what fraction of the tweets are general and what are specific tweets. A very low fractional hyperparameter for a symmetric prior means that we expect the distribution to be very sparse. We do not expect each of the topics to contain all the words in the dictionary, hence we give a low value to the hyperparameter β. α_δ determines the binomial distribution c_t which decides the segment

boundaries for the event transcript, a low value for this beta prior ensures that segment boundaries do not get marked too often. Similar logic is followed for the other two dirichlet priors α_ψ and α_θ which are the segment-topic distribution and the general distribution respectively. These distributions are probably not as sparse as the topic-word distribution, because the number of topics will always be considerably smaller than the number of words, nonetheless we give them fairly lower values because we do not expect a segment to contain too many topics. The exact values assigned to these parameters are mentioned in the evaluation section.

5.3 Results

There are three main things we look to compare - the segmentation of the transcript, the segment distributions and topics discovered and the tweet-segment alignment. The evaluation techniques and baselines are different for each dataset due to the nature of the datasets. With the synthetic data we have the ground truth and we try to make to an objective evaluation of our algorithm, but with real data we use the batch inference algorithm as a baseline and we mainly use user evaluations for measuring the effectiveness of the algorithm.

5.3.1 Evaluation with Denver Debate Data

The parameter values for the number of topics K was 10 and parameter values for the prior dirichlet and beta distributions is as follows

$\alpha_\delta = 0.1$, $\alpha_\gamma = 0.15$, $\alpha_\lambda = [0.3, 0.7]$, $\alpha_\theta = 0.1$, $\alpha_\psi = 0.2$, $\beta = 0.01$

The system was run with the whole event transcript and 15,000 tweets with timestamps between October 3rd 9:00 pm to October 3rd 10:30 pm EDT.

Evaluating the transcript segmentation. We first measure the quality and effectiveness of online ET-LDA for the segmentation of the transcript using the two inference algorithms. A visualization of the results of the event segmentation are shown in Figure 5.1.

We compare the segmentation we have obtained with the hand segmented transcript. P_k measure elaborated in [4], is the probability that a randomly chosen pair of words from the event transcript will be incorrectly separated by a hypothesized segment boundary. Hence, a lower P_k indicative of better agreement with the hand-segmentation hence more accurate.

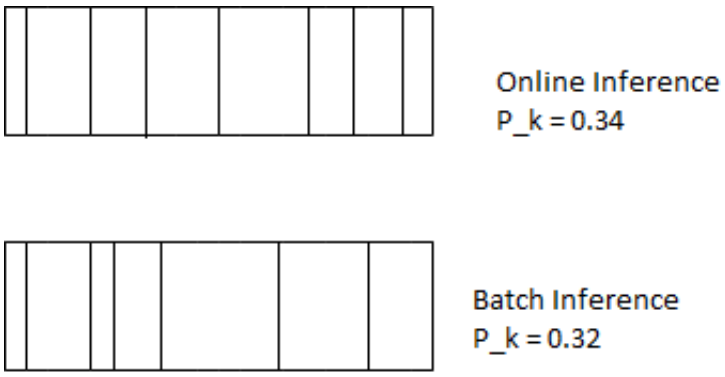

Figure 5.1: Segmentation Boundaries of Denver Debate Transcript(A Lower P_k Values Indicates Better Agreement)

The ground truth against which the P_k values are measured is the hand segmented transcript. The results show that our online inference algorithm has performance comparable to the one achieved by the batch inference.

Topic Distributions. Next we study the topics extracted from the transcript-tweet corpus. The topics extracted from the online inference and baseline, batch inference, are shown in Table 5.1 and 5.2. The tables show the top words for the different topics from the online and batch inference respectively. Each of the topics have been labeled with a name which we believe captures the theme of the top words of that topic. The tables also shows two topics for each of the segments obtained from the segment-topic distribution of the two inference procedures. Table 5.3 shows the user evaluation of the segment-topic distributions obtained.

The user evaluation was done with six graduate students, not associated with the project.

The users were given a copy of the transcript, its segment boundaries, a brief summary of each of the segments, the inferred topic distributions of the segments and top five tweets that aligned with the respective segments. The users were asked to rate the topic relevance of each of the segments and the tweet-segment alignment on a scale of 1-5 with 5 being the best.

Table 5.1: Online Inference

Employment	billion, money, medicaid, programs, approach, care, forward, dollar, young, jobs
Obamacare	care, insurance, the, health, plan, people, obamacare, fact, board
Economy	businesses, deductions, million, money, problem, investments, work, trillion, exemptions
Government	government, know, law, policy, people
Debate	thank, closing, presidential, fight, debate, share, future, individuals, saying
Social Security	medicare, regulation, frank, security, private, social, banks, seniors, need
Regulations	law, people, two, taxes, funding
Tax	middle, tax, we, income, will, class, business, small, top, families, is, things, jobs, america
Defence	war,Iraq, strength

S1	Debate, Government
S2	Employment, Economy
S3	Taxes, Regulation
S4	Economy, Regulations
S5	Social Security, Obamacare
S6	Government, Regulations
S7	Government, Defense
S8	Debate, Government

Table 5.2: Batch Inference

Banking/Regulations	budget, cut, companies, employ, training, employees
Taxes	revenue, trillion, medicaid, taxes, spending, states, raise, corporate, deficit, overseas, dollar, debt, breaks
Economy	percent, work, time, state, money, country, businesses, job, pay
Government	government, know, law, policy, people
Employment	tax, small, jobs, taxes, deficit, energy, trillion, income, families, rates, deductions, business, rate, growth
Obamacare	insurance, health, care, obamacare, medicare
Election/Debate	middle, america, top, states, denver, health, american, college, presidential, university, woman, trickle
Social Security	regulation, that, social, security, banks, health, private, cost, seniors, current, insurance, repeal, retirees, voucher
Defence	Military, elected, wind, american, incomes, party, opportunity, promised
Education	government, education, schools, role, federal, kids, teachers, opportunity, america, elected

S1	Election/Debate
S2	Employment, Economy
S3	Taxes, Banking/Regulation
S4	Economy, Regulations
S5	Social Security, Obamacare
S6	Obamacare, Education
S7	Defense, Government

Table 5.3: Segment-Topic Distribution Evaluation

Inference Procedure	S1	S2	S3	S4	S5	S6	S7
Online Inference	4.83	4.33	3.83	4.0	4.33	3.66	3.5
Batch Inference	4.83	4.16	4.0	4.0	4.33	4.0	4.0

As we can see from the topics extracted and the Likert scale evaluation of the segment-topic distributions, the online inference procedure performs comparably well with respect to the batch inference. The main difference between the two procedures comes up due to the different segment boundaries that get drawn during the inference process. The batch inference is also able to capture extra topics like 'Education' which gets discussed towards the end of the event, but such topics are missed by the online inference.

Tweet-Segment alignment. For each of the segments we retrieve the top 5 tweets that align with it and measure its relevance to the corresponding segment. Likert scale evaluation for the alignment are shown in Table 5.4.

Table 5.4: Segment-Tweet Alignment

Inference Procedure	S2	S3	S4	S5	S6
Online Inference	4.0	3.0	3.5	5.0	3.5
Batch Inference	4.5	4.0	3.16	5.0	3.5

The alignment evaluations are comparable between the two inference procedures. The alignment for the segment that deals with topics like 'Obamacare' and 'Social Security' had a very good performance, this was probably due to the huge popularity of these issues among the general public. We expect a large number of tweets about such topics and users are likely to be more aware and voice specific opinions about such topics, hence we get a very good alignment with such tweets. Also the performance for alignment with the first and last segment was lower compared to other segments because the topics in these

segments were very generic and hence it would be hard to distinguish general tweets and tweets aligning with these segments.

5.3.2 Evolution of the General Topic Discourse

Another interesting aspect we get to observe in online ET-LDA is the evolution of the general topic distribution. As shown in Figure 5.2 we ran online ET-LDA on the whole corpus of twitter data of the Denver debate. As we process the stream of twitter data, the general distribution of the twitter data is continuously evolving, which is reflective of the changing topics in the public discourse of the viewers. We get to observe this evolution of the public discourse, which would not be available in the batch inference process. Also to make such an evaluation we have run ET-LDA on a much larger dataset of twitter data, which spans 2-3 days. The batch inference procedure will have memory constraints and will not be able to handle such huge amounts of data for the inference procedure. The online inference, which reads the data piecewise, will not run into such issues and will be able infer such evolution of discourse over large datasets, spanning the entire timeline of the event.

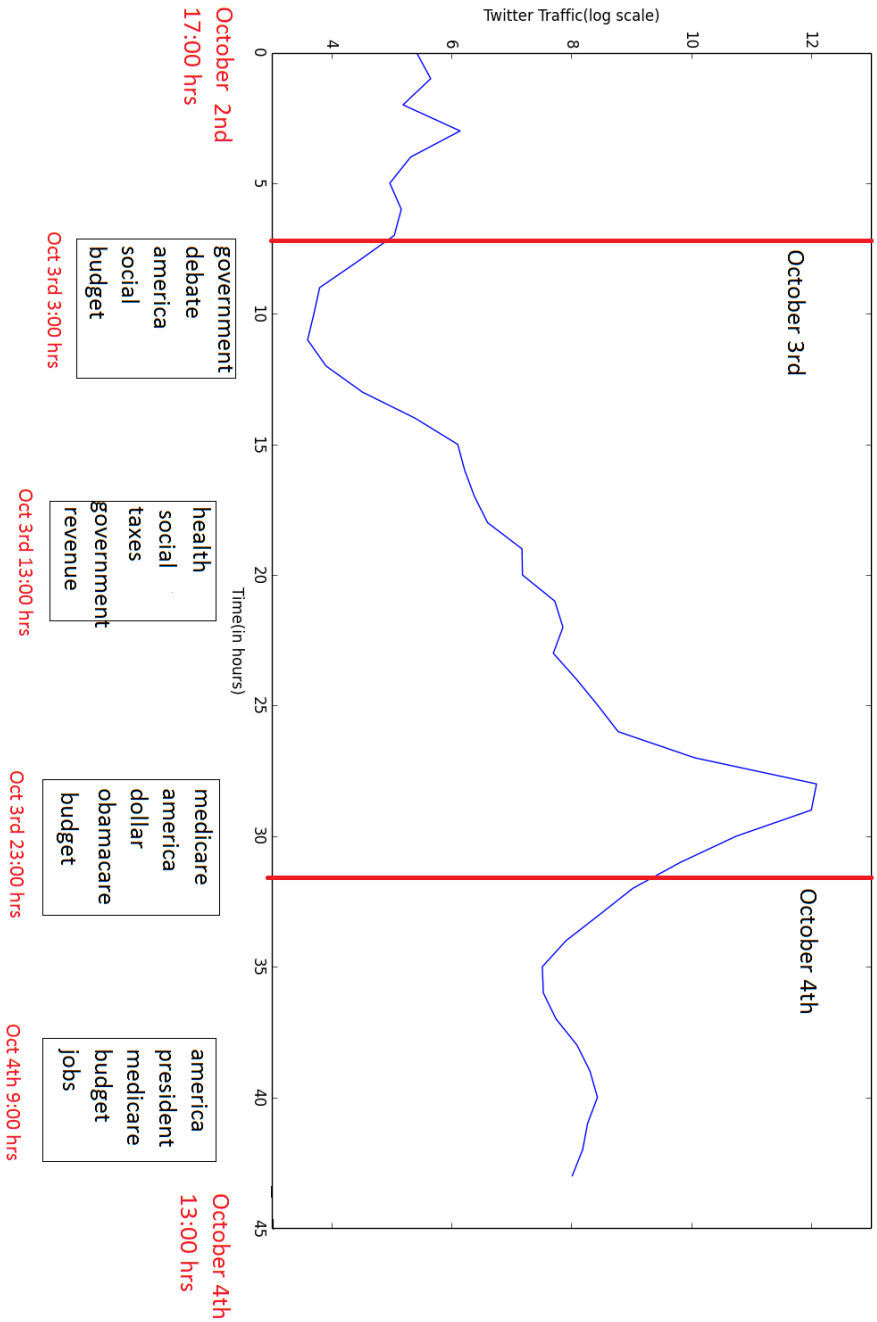

Figure 5.2: Evolution of Public Discourse with Event Timeline

5.3.3 Evaluation with Synthetic Data

Segmentation The segmentation of the online and batch process and their corresponding P_k values is shown in Figure 5.3. The figure shows segmentation obtained by both the online and batch inference, and also the segmented version of the original synthetic data, which is the ground truth in this case. We see the ten segments of the data in the third figure in the diagram, and we also observe how close the segmentation of the online inference procedure comes to the ground truth. Like with Denver debate data, we use P_k values as a metric to determine how close we are able to get to the ground truth. A lower P_k value indicates a better match with the ground truth.

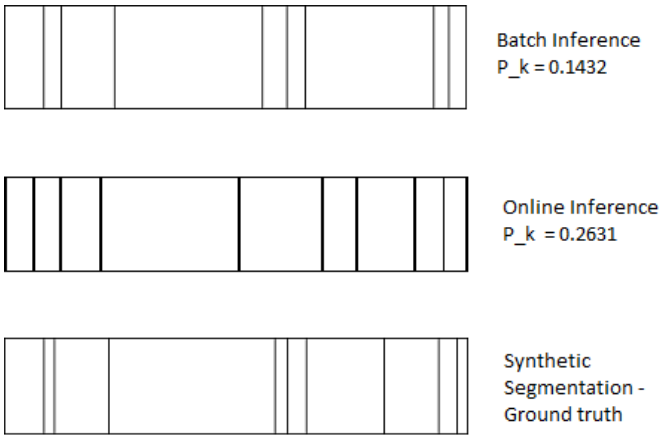

Figure 5.3: Segmentation Boundaries of Synthetic Data(A Lower P_k Values Indicates Better Agreement)

Convergence of Data Likelihood

Figure 5.4 shows the convergence of the log-likelihood over time using both online and batch inference. We took snapshots of the estimated latent distributions of the different variables in the model, at regular intervals of time during the inference procedure and calculated the data likelihood based on these disributions. The time taken by the online inference to converge is lesser than that of the offline process. The graph for the online

inference flattens off much sooner than the offline version, but the likelihood of the batch inference is greater than the online version. This holds true to our expectation that online inference, though faster is less accurate.

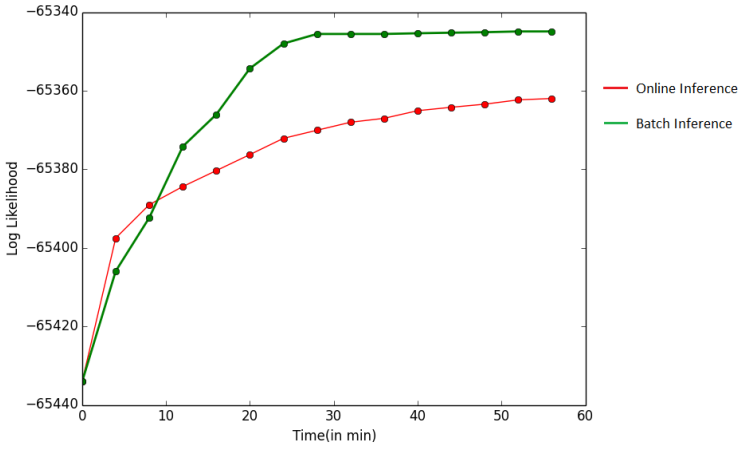

Figure 5.4: Convergence of the Likelihood of the Two Inference Procedures

Apart from a faster convergence the other significant benefit of online inference is scalability. ET-LDA is often run on huge amounts of text and tweets. As the size of the corpus grows it becomes unfeasible to load the whole corpus into memory and run the Gibbs Sampler, the online inference on the other hand does not require you to read the entire twitter corpus into the memory at the same time. At every time epoch, it reads a small subset of the streaming data, that has arrived at that time epoch, hence unlike the batch inference, it never runs out of main memory irrespective of the size of the data.

Chapter 6

CONCLUSION AND FUTURE WORK

In this thesis we introduced an online inference algorithm for ET-LDA, and demonstrated its accuracy and efficiency by evaluating them with both Denver debate and synthetic datasets. We also made a comparison of the runtime and space complexity of the two inference algorithms and found that the online inference algorithm is faster of the two algorithms, and at the same time does not compromise too much on the accuracy of the latent variables. The estimation of the parameters for the graphical model, i.e. the number of topics, determining the dirichlet and beta priors for the different distributions and determining the number of tweets to be included in the rejuvenation phase of the online inference algorithm was accomplished by trial and error. We also found that the online inference algorithm is scalable as it does not have any memory constraints like the batch inference procedure. We also observed and compared the convergence of the log-likelihood of the two inference procedures and saw that though the online inference procedure converges quicker, it is not as accurate as the batch inference procedure. By running the online ET-LDA on the entire tweet corpus, which spans over 2 days of twitter data, we also got to observe a change in the topics of the general distribution, which could be seen as the evolution of the public discourse with respect to the timeline of the event.

6.1 ET-LDA using Bayesian Non-parametrics

One of the possible and the most promising directions of future work will be to make the event transcript also stream online, along with the twitter data. With such an experimental setup we will not only get to see the changing nature of the general distribution but also how the specific distributions of the event transcript evolves. To solve such a problem we

will have to use non-parametric Bayesian methods, where we will not give the number of topics beforehand as a parameter to the model, rather the graphical model and the inference algorithm will automatically converge to the right number of topics. Such a model will very clearly exhibit the evolution of specific topics in the transcript segments. We will not only get to see to see the changing trends in the specific topics but also observe how the number of topics in the event evolves over the duration of the event.

REFERENCES

[1] Twitter statistics, https://about.twitter.com/company.

[2] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[3] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[4] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine learning*, 34(1-3):177–210, 1999.

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

[6] Kevin R Canini, Lei Shi, and Thomas L Griffiths. Online inference of topics with latent dirichlet allocation. In *International conference on artificial intelligence and statistics*, pages 65–72, 2009.

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

[8] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, pages 856–864, 2010.

[9] Yuheng Hu, Ajita John, Fei Wang, and Subbarao Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI*, volume 12, pages 59–65, 2012.

[10] Philip Resnik and Eric Hardisty. Gibbs sampling for the uninitiated, 2009.

[11] Mehran Sahami and Timothy D Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386. AcM, 2006.

[12] Alexander Shraer, Maxim Gurevich, Marcus Fontoura, and Vanja Josifovski. Top-k publish-subscribe for social annotation of news. *Proceedings of the VLDB Endowment*, 6(6):385–396, 2013.

[13] Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198, 2014.