KNOWLEDGE MEDIA

KMi

INSTITUTE

Bayesian Models for Sentence-Level Subjectivity Detection

Technical Report kmi-10-02
June 2010

Yulan He

The Open University

Bayesian Models for Sentence-Level Subjectivity Detection

Yulan He
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes MK7 6AA, UK
y.he@open.ac.uk

Abstract

This paper proposes subjLDA for sentence-level subjectivity detection by modifying the latent Dirichlet allocation (LDA) model through adding an additional layer to model sentence-level subjectivity labels. A variant, called joint-subjLDA, has also been described. The model inference and parameter estimation algorithms, and Gibbs sampling procedure are presented.

1 Introduction

Subjectivity detection automatically identifies whether a give piece of text express sentiment or opinion. Such a task of separating subjective and objective sentences is often viewed as a text classification problem where a classifier is trained from an annotated corpus to perform subjectivity detection.

An early work by Riloff and Wiebe [1] focused on a bootstrapping method to learn subjectivity classifiers from a collection of un-annotated texts. They started with high-precision subjectivity classifiers which automatically identify subjective and objective sentences in un-annotated texts. The subjective extraction patterns were learned from syntactic structure output by Sundance shallow dependency parser [2] from the automatically labeled hight confidence texts. The learned patterns were then used to automatically identify more subjective sentences, which enlarged the training set, and the entire process were then be bootstrapped. As the subjective extraction patterns are based on syntactic structure, they are more flexible than single words or n-grams.

Wilton and Raaijmakers [3] compared the performance of classifiers trained using word n-grams, character n-grams, and phoneme n-grams for recognizing subjective utterances in multiparty conversation. They found that the character n-grams from the reference transcriptions gave the best results out of all the experiments, significantly outperforming word n-grams in terms of subjective recall and F1 score. Raaijmakers *et. al* [4] extended the work in [3] by further analyzing the performance of detecting subjectivity in meeting speech by combining a variety of multimodal features including additional prosodic features. They found that the combination of all features gave the best performance and prosodic features were less useful in discriminating between positive and negative utterances.

More recently, Murray and Carenini [5] proposed to learn subjective patterns from both labeled and unlabeled data using n-gram word sequences with varying level of lexical instantiation. Their approach for learning subjective patterns is similar to [4] which relies on n-grams, but goes beyond fixed sequences of words by varying levels of lexical instantiation as in [1].

Instead of learning subjective extraction patterns or exploring various n-gram features for subjectivity detection, we view the problem as generative model learning and propose a subjLDA Model for sentence-level subjectivity detection. In this model, the generative process involves subjectivity labels for sentences (whether the sentence expresses opinions as being subjective, or states facts as being objective), sentiment label for each word in the sentence (either positive, negative, or neutral), and finally the words in the sentences. If a sentence is subjective ($s_{d,m} = 1$), the words in this sentence can either bear positive polarity ($l_i = 1$), or negative polarity ($l_i = 2$), or are neutral ($l_i = 0$). However, if a sentence is objective ($s_{d,m} = 0$), the words in the sentence are assumed all neutral and they will only be sampled from the word distribution defined for neutral sentiment label φ_0.

The rest of the paper is organized as follows. Section 2 presents the subjLDA model and its inference and parameter estimation algorithms. Section 3 studies a variant of subjLDA called joint-subjLDA. Finally, Section 4 concludes the paper.

Table 1: Notations used in the paper.

Symbol	*Description*
D	number of documents in the collection
K	number of subjectivity labels
M_d	number of sentences in document d
$N_{d,m}$	number of words in sentence m of document d
V	number of unique words
γ	$K-$vector of priors for subjectivity labels
α	matrix of $K \times 3$ dimension, row k represents the mixing proportion of sentiment labels in subjectivity label k
β	$V-$vector of priors for the word distribution conditioned on sentiment labels
π_d	parameter notation for the subjectivity label mixture proportion for document d. $\boldsymbol{\pi} = \{\pi_d\}_{d=1}^{D}$ ($D \times K$ matrix)
$\theta_{d,m}$	mixture proportion of word-sentiments in document d sentence m
ϕ_j	parameter notation for the mixture component for sentiment label j. $\boldsymbol{\phi} = \{\phi_j\}_{j=1}^{3}$ ($3 \times V$ matrix)
$s_{d,m}$	the subjectivity label associated with sentence m in document d
$l_{d,m,t}$	the sentiment label for word t in sentence m of document d
$w_{d,m,t}$	the word t in sentence m of document d
$N_{d,k}$	number of sentences in the dth document assigned to the subjectivity label k
$C_{d,m,j}$	number of words in document d sentence m assigned to sentiment label j
$Y_{j,r}$	number of times the word r assigned to sentiment label j

2 subjLDA

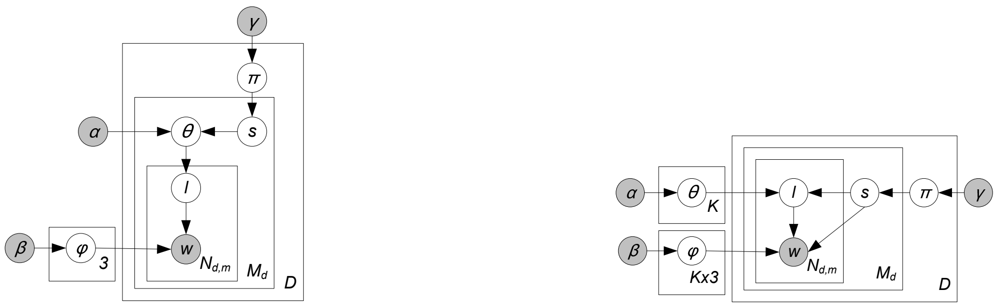
We propose a subjLDA model for sentence-level subjectivity detection. In this model, the generative process involves subjectivity labels for sentences (whether the sentence expresses opinions as being subjective positive, subjective negative, or states facts as being objective), sentiment label for each word in the sentence (either positive, negative, or neutral), and finally the words in the sentences. If a sentence is objective, the words in the sentence are assumed all neutral and they will only be sampled from the word distribution defined for neutral sentiment label.

The generative process which corresponds to the hierarchical Bayesian model is shown in Figure 1(a). A glossary of notations used in the paper is given in Table 1. The generative model is as follows:

- Choose distributions $\varphi \sim Dir(\beta)$.
- For each document $d \in [1, D]$, choose distributions $\pi_d \sim Dir(\gamma)$.
- For each sentence m in document d, $m \in [1, M_D]$,
  - Sample a subjectivity label $s_{d,m} \sim Multinomial(\pi_d)$,
  - Choose a distribution $\theta_{d,m} \sim Dir(\alpha, s_{d,m})$,
  - For each of the $N_{d,m}$ word position w_t,
    * Choose a sentiment label $l_t \sim Multinomial(\theta_{s_{d,m}})$,
    * Choose a word $w_t \sim Multinomial(\varphi_{l_t})$.

The total probability of the model is

$$P(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\pi}; \alpha, \beta, \gamma) = \prod_{j=1}^{3} P(\varphi_j; \beta) \prod_{d=1}^{D} P(\pi_d; \gamma) \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) P(\theta_{d,m}|\alpha, s_{d,m}) \prod_{t=1}^{N_{d,m}} P(l_{d,m,t}|\theta_{d,m}) P(w_{d,m,t}|\varphi_{l_{d,m,t}}) \quad (1)$$



(a) subjLDA

(b) joint-subjLDA

Figure 1: Bayesian models for sentence-level subjectivity detection.

where the bold-font variables denote the vectors.

2.1 Inference and Parameter Estimation

We can integrate out $\boldsymbol{\pi}, \boldsymbol{\theta}$, and $\boldsymbol{\varphi}$.

$$
\begin{aligned}
P(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{s}; \alpha, \beta, \gamma) = & \int_{\boldsymbol{\pi}} \prod_{d=1}^{D} P(\pi_d; \gamma) \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) d\boldsymbol{\pi} \\
& \int_{\boldsymbol{\theta}} \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(\theta_{d,m}|\alpha, s_{d,m}) P(l_{d,m,t}|\theta_{d,m}) d\boldsymbol{\theta} \\
& \int_{\boldsymbol{\varphi}} \prod_{j=1}^{3} P(\varphi_j; \beta) \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(w_{d,m,t}|\varphi_{l_{d,m,t}}) d\boldsymbol{\varphi}
\end{aligned} \tag{2}
$$

First, we focus on $\boldsymbol{\pi}$

$$
\begin{aligned}
\int_{\boldsymbol{\pi}} \prod_{d=1}^{D} P(\pi_d; \gamma) \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) d\boldsymbol{\pi} &= \prod_{d=1}^{D} \int_{\pi_d} P(\pi_d; \gamma) \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) d\pi_d \\
&= \prod_{d=1}^{D} \int_{\pi_d} \frac{\Gamma(\sum_{k=1}^{K} \gamma_k)}{\prod_{k=1}^{K} \Gamma(\gamma_k)} \prod_{k=1}^{K} \pi_{d,k}^{\gamma_k - 1} \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) d\pi_d \\
&= \prod_{d=1}^{D} \int_{\pi_d} \frac{\Gamma(\sum_{k=1}^{K} \gamma_k)}{\prod_{k=1}^{K} \Gamma(\gamma_k)} \prod_{k=1}^{K} \pi_{d,k}^{N_{d,k} + \gamma_k - 1} d\pi_d \\
&= \prod_{d=1}^{D} \frac{\Gamma(\sum_{k=1}^{K} \gamma_k)}{\prod_{k=1}^{K} \Gamma(\gamma_k)} \frac{\prod_{k=1}^{K} \Gamma(N_{d,k} + \gamma_k)}{\Gamma(\sum_{k=1}^{K} N_{d,k} + \gamma_k)}
\end{aligned} \tag{3}
$$

where $N_{d,k}$ denotes the number of sentences in the dth document assigned to the subjectivity label k.

Similarly, for $\boldsymbol{\theta}$

$$
\begin{aligned}
\int_{\boldsymbol{\theta}} \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(l_{d,m,t}|\theta_{d,m}) d\boldsymbol{\theta} &= \prod_{d=1}^{D} \prod_{m=1}^{M_d} \int_{\theta_{d,m}} \prod_{t=1}^{N_{d,m}} P(\theta_{d,m}|\alpha, s_{d,m}) P(l_{d,m,t}|\theta_{d,m}) d\theta_{d,m} \\
&= \prod_{d=1}^{D} \prod_{m=1}^{M_d} \int_{\theta_{d,m}} \frac{\Gamma(\sum_{j=1}^{3} \alpha_{s_{d,m},j})}{\prod_{j=1}^{3} \Gamma(\alpha_{s_{d,m},j})} \prod_{j=1}^{3} \theta_{d,m,j}^{\alpha_{s_{d,m},j} - 1} \prod_{j=1}^{3} \theta_{d,m,j}^{C_{d,m,j}} d\theta_{d,m} \\
&= \prod_{d=1}^{D} \prod_{m=1}^{M_d} \frac{\Gamma(\sum_{j=1}^{3} \alpha_{s_{d,m},j})}{\prod_{j=1}^{3} \Gamma(\alpha_{s_{d,m},j})} \frac{\prod_{j=1}^{3} \Gamma(C_{d,m,j} + \alpha_{s_{d,m},j})}{\Gamma(\sum_{j=1}^{3} C_{d,m,j} + \alpha_{s_{d,m},j})}
\end{aligned} \tag{4}
$$

where $C_{d,m,j}$ denotes the number of words in sentences in document d sentence m assigned to sentiment label j.

Finally, for $\boldsymbol{\varphi}$

$$\begin{aligned}
\int_{\boldsymbol{\varphi}} \prod_{j=1}^{3} P(\varphi_j;\beta) \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(w_{d,m,t}|\varphi_{l_{d,m,t}}) d\boldsymbol{\varphi} &= \prod_{j=1}^{3} \int_{\varphi_j} P(\varphi_j;\beta) \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(w_{d,m,t}|\varphi_{l_{d,m,t}}) d\varphi_j \\
&= \prod_{j=1}^{3} \int_{\varphi_j} \frac{\Gamma(\sum_{r=1}^{V}\beta_r)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \prod_{r=1}^{V} \varphi_{j,r}^{\beta_r-1} \prod_{r=1}^{V} \varphi_{j,r}^{Y_{j,r}} d\varphi_j \\
&= \prod_{j=1}^{3} \frac{\Gamma(\sum_{r=1}^{V}\beta_r)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \frac{\prod_{r=1}^{V}\Gamma(Y_{j,r}+\beta_r)}{\Gamma(\sum_{r=1}^{V}Y_{j,r}+\beta_r)} \qquad (5)
\end{aligned}$$

where $Y_{j,r}$ denotes the number of times the word r assigned to sentiment label j.

The final equation with $\boldsymbol{\pi}, \boldsymbol{\theta}$, and $\boldsymbol{\varphi}$ integrated out is:

$$\begin{aligned}
P(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{s}; \alpha, \beta, \gamma) = &\prod_{d=1}^{D} \frac{\Gamma(\sum_{k=1}^{K}\gamma_k)}{\prod_{k=1}^{K}\Gamma(\gamma_k)} \frac{\prod_{k=1}^{K}\Gamma(N_{d,k}+\gamma_k)}{\Gamma(\sum_{k=1}^{K}N_{d,k}+\gamma_k)} \\
&\prod_{d=1}^{D} \prod_{m=1}^{M_d} \frac{\Gamma(\sum_{j=1}^{3}\alpha_{s_{d,m},j})}{\prod_{j=1}^{3}\Gamma(\alpha_{s_{d,m},j})} \frac{\prod_{j=1}^{3}\Gamma(C_{d,m,j}+\alpha_{s_{d,m},j})}{\Gamma(\sum_{j=1}^{3}C_{d,m,j}+\alpha_{s_{d,m},j})} \\
&\prod_{j=1}^{3} \frac{\Gamma(\sum_{r=1}^{V}\beta_r)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \frac{\prod_{r=1}^{V}\Gamma(Y_{j,r}+\beta_r)}{\Gamma(\sum_{r=1}^{V}Y_{j,r}+\beta_r)} \qquad (6)
\end{aligned}$$

Gibbs sampling will sequentially sampling each variable of interest, $s_{d,m}$ and $l_{d,m,t}$ here, from the distribution over that variable given the current values of all other variables and the data. Letting the index $x = (d, m)$ and the subscript $-x$ denote a quantity that excludes counts in sentence m of document d, the conditional posterior for s_x is:

$$\begin{aligned}
P(s_x = k|\boldsymbol{s}_{-x}, \boldsymbol{l}, \boldsymbol{w}, \alpha, \beta, \gamma) &= \frac{P(s_x, l_x, w_x|\boldsymbol{s}_{-x}, \boldsymbol{l}_{-x}, \boldsymbol{w}_{-x}, \alpha, \beta, \gamma)}{P(l_x, w_x|\boldsymbol{s}_{-x}, \boldsymbol{l}_{-x}, \boldsymbol{w}_{-x}, \alpha, \beta, \gamma)} \\
&\propto \frac{P(\boldsymbol{s}, \boldsymbol{l}, \boldsymbol{w}, \alpha, \beta, \gamma)}{P(\boldsymbol{s}_{-x}, \boldsymbol{l}_{-x}, \boldsymbol{w}_{-x}, \alpha, \beta, \gamma)} \\
&\propto \frac{\{N_{d,k}\}_{-x} + \gamma}{\{M_d\}_{-x} + K\gamma} \cdot \frac{\prod_{j=1}^{3}\prod_{b=0}^{C_{d,m,j}-1}(b+\alpha_{s_x,j})}{\prod_{b=0}^{C_{d,m}-1}(b+\sum_{j=1}^{3}\alpha_{s_x,j})} \qquad (7)
\end{aligned}$$

where $N_{d,k}$ denotes the frequency of sentences assigned to subjectivity label k in document d, M_d is the total number of sentences in document d, $C_{d,m,j}$ is the frequency of words with sentiment label j in document d sentence m, $C_{d,m}$ is the total number of words in document d sentence m.

Letting the index $y = (d, m, t)$ denote t^{th} word in sentence m of document d and the subscript $-y$ denote a quantity that excludes data from t^{th} word position, the conditional posterior for l_y is:

$$\begin{aligned}
P(l_y = j|\boldsymbol{s}, \boldsymbol{l}_{-y}, \boldsymbol{w}, \alpha, \beta, \gamma) &= \frac{P(l_y = j, w_y|\boldsymbol{s}, \boldsymbol{l}_{-y}, \boldsymbol{w}_{-y}, \alpha, \beta, \gamma)}{P(w_y|\boldsymbol{s}, \boldsymbol{l}_{-y}, \boldsymbol{w}_{-y}, \alpha, \beta, \gamma)} \\
&\propto \frac{P(\boldsymbol{s}, \boldsymbol{l}, \boldsymbol{w}, \alpha, \beta, \gamma)}{P(\boldsymbol{s}, \boldsymbol{l}_{-y}, \boldsymbol{w}_{-y}, \alpha, \beta, \gamma)} \\
&= \frac{\{C_{d,m,j}\}_{-y} + \alpha s_{d,m}, j}{\{C_{d,m}\}_{-y} + \sum_{j=1}^{3}\alpha_{s_{d,m},j}} \cdot \frac{\{Y_{j,w_t}\}_{-y} + \beta}{\{Y_j\}_{-y} + V\beta}. \qquad (8)
\end{aligned}$$

where Y_{j,w_t} denotes the frequency of word w_t with sentiment label j in the document collection, Y_j denotes the total number of words with sentiment label j in the document collection.

Equations 7-8 are the conditional probabilities derived by marginalizing out the random variables π, θ, and φ. A sample obtained from the Markov chain can be used to approximate the distribution over subjectivity label for sentence:

$$\pi_{d,k} = \frac{N_{d,k} + \gamma}{N_d + K\gamma} \tag{9}$$

The approximated predictive distribution over words for sentimental label is:

$$\theta_{d,m,j} = \frac{C_{d,m,j} + \alpha_{s_{d,m},j}}{C_{d,m} + \sum_{j=1}^{3} \alpha_{s_{d,m},j}} \tag{10}$$

Finally, the approximated predictive distribution of words in sentiment labels:

$$\varphi_{j,r} = \frac{Y_{j,r} + \beta}{Y_j + V\beta} \tag{11}$$

The Gibbs sampling procedure is given in Algorithm 1.

Algorithm 1 Gibbs sampling procedure for subjLDA.

Input: α, β, γ, Corpus
Output: sentiment assignment for all words and subjectivity label for sentences

Initialize all count variables $N_{d,k}$, N_d, $C_{d,m,j}$, $C_{d,m}$, $Y_{j,r}$, Y_j.
Randomize the order of documents in the corpus, the order of sentences in each document, and the order of words in each sentence.
for $i = 1$ to max Gibbs sampling iterations **do**
    for all documents $d \in [1, D]$ **do**
        for all sentences $m \in [1, M_d]$ **do**
            Exclude sentence m and its assigned subjectivity label k from variables $N_{d,k}$, N_d
            sample a new subjectivity label $\tilde{s}_{d,m}$ for sentence m using Equation 7
            Update variables $N_{d,k}$, N_d using the new subjectivity label $\tilde{s}_{d,m}$
            for all words $t \in [1, N_{d,m}]$ **do**
                Exclude word t and its assigned sentiment label l from variables $C_{d,m,j}$, $C_{d,m}$, $Y_{j,r}$, Y_j
                Sample a new sentiment label $\tilde{l}_{d,m,t}$ using Equation 8
                Update variables $C_{d,m,j}$, $C_{d,m}$, $Y_{j,r}$, Y_j using the new sentiment label $\tilde{l}_{d,m,t}$
            end for
        end for
    end for
    if converged and I sampling iterations since last read out **then**
        Update the matrix $\boldsymbol{\varphi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\pi}$ with new sampling results
    end if
end for

---

3 joint-subjLDA

It is also possible to slightly change the topology of the original subjLDA and we get a model as depicted in Figure 1(b), which is called joint-subjLDA. The generative model of joint-subjLDA is as follows:

- Choose distributions $\varphi \sim Dir(\beta)$.
- Choose a distribution $\theta \sim Dir(\alpha)$,
- For each document $d \in [1, D]$, choose distributions $\pi_d \sim Dir(\gamma)$.
- For each sentence m in document d, $m \in [1, M_D]$,
  - Sample a subjectivity label $s_{d,m} \sim Multinomial(\pi_d)$,

– For each of the $N_{d,m}$ word position w_t,
  * Choose a sentiment label $l_t \sim Multinomial(\theta_{s_{d,m}})$,
  * Choose a word $w_t \sim Multinomial(\varphi^{l_t}_{s_{d,m}})$.

The total probability of the model is

$$P(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\pi}; \alpha, \beta, \gamma) = \prod_{k=1}^{K} P(\theta_k; \alpha) \prod_{j=1}^{3} P(\varphi_{k,j}; \beta) \prod_{d=1}^{D} P(\pi_d; \gamma) \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) \prod_{t=1}^{N_{d,m}} P(l_{d,m,t}|\theta_{s_{d,m}}) P(w_{d,m,t}|\varphi^{l_{d,m,t}}_{s_{d,m}}) \quad (12)$$

By integrating out $\boldsymbol{\pi}, \boldsymbol{\theta}$, and $\boldsymbol{\varphi}$

$$P(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{s}; \alpha, \beta, \gamma) = \int_{\boldsymbol{\pi}} \prod_{d=1}^{D} P(\pi_d; \gamma) \prod_{m=1}^{M_d} P(s_{d,m}|\pi_d) d\boldsymbol{\pi} \int_{\boldsymbol{\theta}} \prod_{k=1}^{K} P(\theta_k; \alpha) \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(l_{d,m,t}|\theta_{s_{d,m}}) d\boldsymbol{\theta} \int_{\boldsymbol{\varphi}} \prod_{k=1}^{K} \prod_{j=1}^{3} P(\varphi_{k,j}; \beta) \prod_{d=1}^{D} \prod_{m=1}^{M_d} \prod_{t=1}^{N_{d,m}} P(w_{d,m,t}|\varphi^{l_{d,m,t}}_{s_{d,m}}) d\boldsymbol{\varphi}, \quad (13)$$

we get

$$P(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{s}; \alpha, \beta, \gamma) = \prod_{d=1}^{D} \frac{\Gamma(\sum_{k=1}^{K} \gamma_k)}{\prod_{k=1}^{K} \Gamma(\gamma_k)} \frac{\prod_{k=1}^{K} \Gamma(N_{d,k} + \gamma_k)}{\Gamma(\sum_{k=1}^{K} N_{d,k} + \gamma_k)} \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{3} \alpha_{k,j})}{\prod_{j=1}^{3} \Gamma(\alpha_{k,j})} \frac{\prod_{j=1}^{3} \Gamma(C_{k,j} + \alpha_{k,j})}{\Gamma(\sum_{j=1}^{3} C_{k,j} + \alpha_{k,j})} \prod_{k=1}^{K} \prod_{j=1}^{3} \frac{\Gamma(\sum_{r=1}^{V} \beta_r)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \frac{\prod_{r=1}^{V} \Gamma(C_{k,j,r} + \beta_r)}{\Gamma(\sum_{r=1}^{V} C_{k,j,r} + \beta_r)}. \quad (14)$$

Letting the index $x = (d, m)$ and the subscript $-x$ denote a quantity that excludes counts in sentence m of document d, the conditional posterior for s_x is:

$$P(s_x = k|\boldsymbol{s}_{-x}, \boldsymbol{l}, \boldsymbol{w}, \alpha, \beta, \gamma) \propto \frac{\{N_{d,k}\}_{-x} + \gamma}{\{M_d\}_{-x} + K\gamma} \cdot \prod_{j=1}^{3} \frac{\prod_{b=1}^{C_{k,j}^{(d,m)}} (C_{k,j} - b + \alpha_{k,j})}{\prod_{b=1}^{C_k^{(d,m)}} (C_k - b + \sum_{j=1}^{3} \alpha_{k,j})} \cdot \prod_{j=1}^{3} \prod_{r \in \{W_{d,m,j}\}} \frac{\prod_{b=1}^{C_{k,j,r}^{(d,m)}} (C_{k,j,r} - b + \beta)}{\prod_{b=1}^{C_{k,j}^{(d,m)}} (C_{k,j} - b + V\beta)} \quad (15)$$

where $N_{d,k}$ denotes the frequency of sentences assigned to subjectivity label k in document d, M_d is the total number of sentences in document d, $C_{k,j}$ is the frequency of words with sentiment label j in sentences with subjectivity label k, C_k is the total number of words in sentences with subjectivity label k, $C_{k,j,r}$ is the frequency of word r with sentiment label j in sentences with subjectivity label k, $W_{d,m,j}$ is the word token in document d sentence m with sentiment label j, counts with (d, m) notation denote the count relating to document d, sentence m only.

Letting the index $y = (d, m, t)$ denote t^{th} word in sentence m of document d and the subscript $-y$ denote a quantity that excludes data from t^{th} word position, the conditional posterior for l_y is:

$$P(l_y = j|\boldsymbol{s}, \boldsymbol{l}_{-y}, \boldsymbol{w}, \alpha, \beta, \gamma) \propto \frac{\{C_{s_{d,m},j}\}_{-y} + \alpha_{s_{d,m},j}}{\{C_{s_{d,m}}\}_{-y} + \sum_{j=1}^{3} \alpha_{s_{d,m},j}} \cdot \frac{\{C_{k,j,w_t}\}_{-y} + \beta}{\{C_{k,j}\}_{-y} + V\beta}. \quad (16)$$

where C_{k,j,w_t} denotes the frequency of word w_t with sentiment label j in sentences with subjectivity label k.

Equations 15-16 are the conditional probabilities derived by marginalizing out the random variables π, θ, and φ. A sample obtained from the Markov chain can be used to approximate the distribution over subjectivity label for sentence:

$$\pi_{d,k} = \frac{N_{d,k} + \gamma}{N_d + K\gamma} \tag{17}$$

The approximated predictive distribution over words for sentimental label is:

$$\theta_{k,j} = \frac{C_{k,j} + \alpha_{k,j}}{C_k + \sum_{j=1}^{3} \alpha_{k,j}} \tag{18}$$

Finally, the approximated predictive distribution of words in sentiment labels:

$$\varphi_{k,j,r} = \frac{C_{k,j,r} + \beta}{C_{k,j} + V\beta} \tag{19}$$

4 Conclusions

This paper has proposed a subjLDA model for subjectivity detection. The inference and parameter estimation algorithms have been presented. A variant of subjLDA called joint-subjLDA has also been discussed.

References

[1] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 105–112, 2003.

[2] E. Riloff and W. Phillips. An introduction to the Sundance and AutoSlog Systems. Technical report, School of Computing, University of Utah, 2004.

[3] T. Wilson and S. Raaijmakers. Comparing word, character, and phoneme n-grams for subjective utterance recognition. In *Proceedings of InterSpeech*, pages 1614–1617, 2008.

[4] S. Raaijmakers, K. Truong, and T. Wilson. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 466–474, 2008.

[5] G. Murray and G. Carenini. Predicting Subjectivity in Multimodal Conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1348–1357, 2009.