A Comparative Study of Bayesian Models for Unsupervised Sentiment Detection

Chenghua Lin
School of Engineering,
Computing and Mathematics
University of Exeter
Exeter, EX4 4QF, UK.
cl322@exeter.ac.uk

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

Richard Everson
School of Engineering,
Computing and Mathematics
University of Exeter
Exeter, EX4 4QF, UK.
R.E.Everson@exeter.ac.uk

Abstract

This paper presents a comparative study of three closely related Bayesian models for unsupervised document level sentiment classification, namely, the latent sentiment model (LSM), the joint sentiment-topic (JST) model, and the Reverse-JST model. Extensive experiments have been conducted on two corpora, the movie review dataset and the multi-domain sentiment dataset. It has been found that while all the three models achieve either better or comparable performance on these two corpora when compared to the existing unsupervised sentiment classification approaches, both JST and Reverse-JST are able to extract sentiment-oriented topics. In addition, Reverse-JST always performs worse than JST suggesting that the JST model is more appropriate for joint sentiment topic detection.

1 Introduction

With the explosion of web 2.0, various types of social media such as blogs, discussion forums and peer-to-peer networks present a wealth of information that can be very helpful in assessing the general public's sentiments and opinions towards products and services. Recent surveys have revealed that opinion-rich resources like online reviews are having greater economic impact on both consumers and companies compared to the traditional media (Pang and Lee, 2008). Driven by the demand of gleaning insights of such great amounts of user-generated data, work on new methodologies for automated sentiment analysis has bloomed splendidly.

Compared to the traditional topic-based text classification, sentiment classification is deemed to be more challenging as sentiment is often embodied in subtle linguistic mechanisms such as the use of sarcasm or incorporated with highly domain-specific information. Although the task of identifying the overall sentiment polarity of a document has been well studied, most of the work is highly domain dependent and favoured in supervised learning (Pang et al., 2002; Pang and Lee, 2004; Whitelaw et al., 2005; Kennedy and Inkpen, 2006; McDonald et al., 2007), requiring annotated corpora for every possible domain of interest, which is impractical for real applications. Also, it is well-known that sentiment classifiers trained on one domain often fail to produce satisfactory results when shifted to another domain, since sentiment expression can be quite different in different domains (Aue and Gamon, 2005). Moreover, aside from the diversity of genres and large-scale size of Web corpora, user-generated contents evolve rapidly over time, which demands much more efficient algorithms for sentiment analysis than the current approaches can offer. These observations have thus motivated the problem of using unsupervised approaches for domain-independent joint sentiment topic detection.

Some recent research efforts have been made to adapt sentiment classifiers trained on one domain to another domain (Aue and Gamon, 2005; Blitzer et al., 2007; Li and Zong, 2008; Andreevskaia and Bergler, 2008). However, the adaption performance of these lines of work pretty much depends on the distribution similarity between the source and target domain, and considerable effort is still required to obtain labelled data for training.

Intuitively, sentiment polarities are dependent on contextual information, such as topics or domains. In this regard, some recent work (Mei et al., 2007; Titov and McDonald, 2008a) has tried to model both sentiment and topics. However, these two models either require postprocessing to calculate the positive/negative coverage in a document for polarity identification (Mei et al., 2007) or re-

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 144–152,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

quire some kind of supervised setting in which review text should contain ratings for aspects of interest (Titov and McDonald, 2008a). More recently, Dasgupta and Ng (2009) proposed an unsupervised sentiment classification algorithm by integrating user feedbacks into a spectral clustering algorithm. Features induced for each dimension of spectral clustering can be considered as sentiment-oriented topics. Nevertheless, human judgement of identifying the most important dimensions during spectral clustering is required.

Lin and He (2009) proposed a joint sentiment-topic (JST) model for unsupervised joint sentiment topic detection. They assumed that topics are generated dependent on sentiment distributions and then words are generated conditioned on sentiment-topic pairs. While this is a reasonable design choice, one may argue that the reverse is also true that sentiments may vary according to topics. Thus in this paper, we studied the reverse dependence of the JST model called Reverse-JST, in which sentiments are generated dependent on topic distributions in the modelling process. We also note that, when the topic number is set to 1, both JST and reversed-JST essentially become a simple latent Dirichlet allocation (LDA) model with only S (number of sentiment label) topics, each of which corresponds to a sentiment label. We called it latent sentiment model (LSM) in this paper. Extensive experiments have been conducted on the movie review (MR)[1] (Pang et al., 2002) and multi-domain sentiment (MDS)[2] (Blitzer et al., 2007) datasets to compare the performance of LSM, JST and Reverse-JST. Results show that all these three models are able to give either better or comparable performance compared to the existing unsupervised sentiment classification approaches. In addition, both JST and reverse-JST are able to extract sentiment-oriented topics. Furthermore, the fact that reverse-JST always performs worse than JST suggests that the JST model is more appropriate for joint sentiment topic detection.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes the LSM, JST and Reserver-JST models. Experimental setup and results on the MR and MDS datasets are discussed in Section 4 and 5 respectively. Finally, Section 6 concludes the paper and outlines the future work.

2 Related Work

As opposed to the work (Pang et al., 2002; Pang and Lee, 2004; Whitelaw et al., 2005; Kennedy and Inkpen, 2006) that only focused on sentiment classification in one particular domain, recent research attempts have been made to address the problem of sentiment classification across domains. Aue and Gamon (2005) explored various strategies for customizing sentiment classifiers to new domains, where the training is based on a small number of labelled examples and large amounts of unlabelled in-domain data. However, their experiments achieved only limited success, with most of the classification accuracy below 80%. In the same vein, some more recent work focused on domain adaption for sentiment classifiers. Blitzer et al. (2007) used the structural correspondence learning (SCL) algorithm with mutual information. Li and Zong (2008) combined multiple single classifiers trained on individual domains using SVMs. However, the adaption performance in (Blitzer et al., 2007) depends on the selection of pivot features that used to link the source and target domains; whereas the approach of Li and Zong (2008) heavily relies on labelled data from all the domains to train the integrated classifier and thus lack the flexibility to adapt the trained classifier to domains where label information is not available.

Recent years have also seen increasing interests in modelling both sentiment and topics simultaneously. The topic-sentiment mixture (TSM) model (Mei et al., 2007) can jointly model sentiment and topics by constructing an extra background component and two additional sentiment subtopics on top of the probabilistic latent semantic indexing (pLSI) (Hofmann, 1999). However, TSM may suffer from the problem of overfitting the data which is known as a deficiency of pLSI, and post-processing is also required in order to calculate the sentiment prediction for a document. The multi-aspect sentiment (MAS) model (Titov and McDonald, 2008a), which is extended from the multi-grain latent Dirichlet allocation (MG-LDA) model (Titov and McDonald, 2008b), allows sentiment text aggregation for sentiment summary of each rating aspect extracted from MG-LDA. One drawback of MAS is that it requires that every aspect is rated at least in some documents, which

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data

[2] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html

is practically infeasible. More recently, Dasgupta and Ng (2009) proposed an unsupervised sentiment classification algorithm where user feedbacks are provided on the spectral clustering process in an interactive manner to ensure that text are clustered along the sentiment dimension. Features induced for each dimension of spectral clustering can be considered as sentiment-oriented topics. Nevertheless, human judgement of identifying the most important dimensions during spectral clustering is required.

Among various efforts for improving sentiment detection accuracy, one direction is to incorporate prior information or subjectivity lexicon (i.e., words bearing positive or negative sentiment) into the sentiment model. Such sentiment lexicons can be acquired from domain-independent sources in many different ways, from manually built appraisal groups (Whitelaw et al., 2005), to semi-automatically (Abbasi et al., 2008) and fully automatically (Kaji and Kitsuregawa, 2006) constructed lexicons. When incorporating lexical knowledge as prior information into a sentiment-topic model, Andreevskaia and Bergler (2008) integrated the lexicon-based and corpus-based approaches for sentence-level sentiment annotation across different domains; Li et al. (2009) employed lexical prior knowledge for semi-supervised sentiment classification based on non-negative matrix tri-factorization, where the domain-independent prior knowledge was incorporated in conjunction with domain-dependent unlabelled data and a few labelled documents. However, this approach performed worse than the JST model on the movie review data even with 40% labelled documents as will be shown in Section 5.

3 Latent Sentiment-Topic Models

This section describes three closely related Bayesian models for unsupervised sentiment classification, the latent sentiment model (LSM), the joint sentiment-topic (JST) model, and the joint topic sentiment model by reversing the generative process of sentiment and topics in the JST model, called Reverse-JST.

3.1 Latent Sentiment Model (LSM)

The LSM model, as shown in Figure 1(a), can be treated as a special case of LDA where a mixture of only three sentiment labels are modelled, i.e. positive, negative and neutral.

Assuming that we have a total number of S sentiment labels[3]; a corpus with a collection of D documents is denoted by $C = \{d_1, d_2, ..., d_D\}$; each document in the corpus is a sequence of N_d words denoted by $d = (w_1, w_2, ..., w_{N_d})$, and each word in the document is an item from a vocabulary index with V distinct terms denoted by $\{1, 2, ..., V\}$. The procedure of generating a word in LSM starts by firstly choosing a distribution over three sentiment labels for a document. Following that, one picks up a sentiment label from the sentiment label distribution and finally draws a word according to the sentiment label-word distribution.

The joint probability of words and sentiment label assignment in LSM can be factored into two terms:

$$P(\mathbf{w}, \mathbf{l}) = P(\mathbf{w}|\mathbf{l})P(\mathbf{l}|d). \tag{1}$$

Letting the superscript $-t$ denote a quantity that excludes data from the t^{th} position, the conditional posterior for l_t by marginalizing out the random variables φ and π is

$$P(l_t = k|\mathbf{w}, \mathbf{l^{-t}}, \beta, \boldsymbol{\gamma}) \propto \frac{N_{w_t,k}^{-t} + \beta}{N_k^{-t} + V\beta} \cdot \frac{N_{k,d}^{-t} + \gamma_k}{N_d^{-t} + \sum_k \gamma_k}, \tag{2}$$

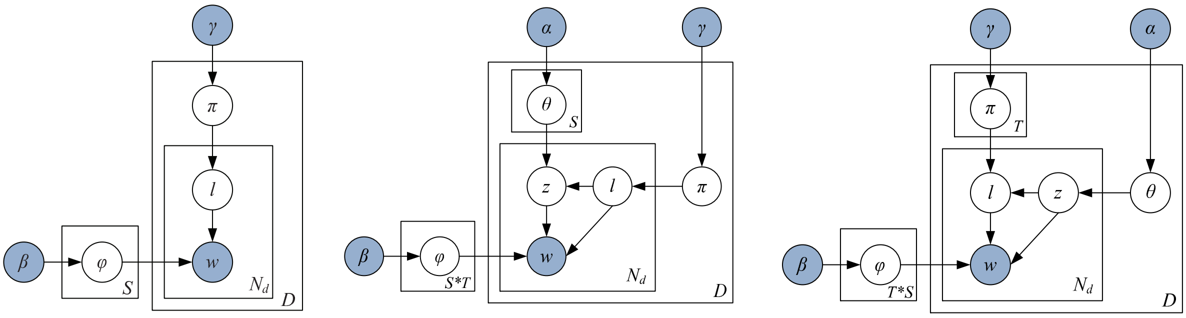
where $N_{w_t,k}$ is the number of times word w_t has associated with sentiment label k; N_k is the the number of times words in the corpus assigned to sentiment label k; $N_{k,d}$ is the number of times sentiment label k has been assigned to some word tokens in document d; N_d is the total number of words in the document collection.

Gibbs sampling is used to estimate the posterior distribution of LSM, as well as the JST and Reverse-JST models that will be discussed in the following two sections.

3.2 Joint Sentiment-Topic Model (JST)

In contrast to LSM that only models document sentiment, the JST model (Lin and He, 2009) can detect sentiment and topic simultaneously, by modelling each document with S (number of sentiment labels) topic-document distributions. It should be noted that when the topic number is set to 1, JST effectively becomes the LSM model with only three topics corresponding to each of the

[3] For all the three models, i.e., LSM, JST and Reverse-JST, we set the sentiment label number S to 3 representing the positive, negative and neutral polarities, respectively.



(a) (b) (c)

Figure 1: (a) LSM model; (b) JST model; (c) Reverse-JST model.

three sentiment labels. Let T be the total number of topics, the procedure of generating a word w_i according to the graphical model shown in Figure 1(b) is:

- For each document d, choose a distribution $\pi_d \sim \text{Dir}(\gamma)$.
- For each sentiment label l of document d, choose a distribution $\theta_{d,l} \sim \text{Dir}(\alpha)$.
- For each word w_i in document d
  - choose a sentiment label $l_i \sim$ Multinomial(π_d),
  - choose a topic $z_i \sim$ Multinomial(θ_{d,l_i}),
  - choose a word w_i from $\varphi^{l_i}_{z_i}$, a Multinomial distribution over words conditioned on topic z_i and sentiment label l_i.

In JST, the joint probability of words and topic-sentiment label assignments can be factored into three terms:

$$P(\mathbf{w}, \mathbf{z}, \mathbf{l}) = P(\mathbf{w}|\mathbf{z}, \mathbf{l})P(\mathbf{z}|\mathbf{l}, d)P(\mathbf{l}|d). \quad (3)$$

The conditional posterior for z_t and l_t can be obtained by marginalizing out the random variables φ, θ, and π:

$$P(z_t = j, l_t = k|\mathbf{w}, \mathbf{z}^{-\mathbf{t}}, \mathbf{l}^{-\mathbf{t}}, \alpha, \beta, \boldsymbol{\gamma}) \propto$$
$$\frac{N^{-t}_{w_t,j,k} + \beta}{N^{-t}_{j,k} + V\beta} \cdot \frac{N^{-t}_{j,k,d} + \alpha}{N^{-t}_{k,d} + T\alpha} \cdot \frac{N^{-t}_{k,d} + \gamma_k}{N^{-t}_{d} + \sum_k \gamma_k}, \quad (4)$$

where $N_{w_t,j,k}$ is the number of times word w_t appeared in topic j and with sentiment label k; $N_{j,k}$ is the number of times words assigned to topic j and sentiment label k, $N_{k,j,d}$ is the number of times a word from document d has been associated with topic j and sentiment label k; $N_{k,d}$ is the number of times sentiment label k has been assigned to some word tokens in document d.

3.3 Reverse Joint Sentiment-Topic Model (Reverse-JST)

We also studied a variant of the JST model, called Reverse-JST. As opposed to JST in which topic generation is conditioned on sentiment labels, sentiment label generation in Reverse-JST is dependent on topics. As shown in Figure 1(c), Reverse-JST is effectively a four-layer hierarchical Bayesian model, where topics are associated with documents, under which sentiment labels are associated with topics and words are associated with both topics and sentiment labels.

The procedure of generating a word w_i in Reverse-JST is shown below:

- For each document d, choose a distribution $\theta_d \sim \text{Dir}(\alpha)$.
- For each topic z of document d, choose a distribution $\pi_{d,z} \sim \text{Dir}(\gamma)$.
- For each word w_i in document d
  - choose a topic $z_i \sim$ Multinomial(θ_d),
  - choose a sentiment label $l_i \sim$ Multinomial(π_{d,z_i}),
  - choose a word w_i from $\varphi^{l_i}_{z_i}$, a multinomial distribution over words conditioned on the topic z_i and sentiment label l_i.

Analogy to JST, in Reverse-JST the joint probability of words and the topic-sentiment label assignments can be factored into the following three terms:

$$P(\mathbf{w}, \mathbf{l}, \mathbf{z}) = P(\mathbf{w}|\mathbf{l}, \mathbf{z})P(\mathbf{l}|\mathbf{z}, d)P(\mathbf{z}|d), \quad (5)$$

and the conditional posterior for z_t and l_t can be derived by integrating out the random variables φ,

θ, and π, yielding

$$P(z_t = j, l_t = k | \mathbf{w}, \mathbf{z}_{-t}, \mathbf{l}_{-t}, \alpha, \beta, \boldsymbol{\gamma}) \propto \frac{N^{-t}_{w_t,j,k} + \beta}{N^{-t}_{j,k} + V\beta} \cdot \frac{N^{-t}_{k,j,d} + \gamma_k}{N^{-t}_{j,d} + \sum_k \gamma_k} \cdot \frac{N^{-t}_{j,d} + \alpha}{N^{-t}_d + T\alpha}. \quad (6)$$

It it noted that most of the terms in the Reverse-JST posterior is identical to the posterior of JST in Equation 4, except that $N_{j,d}$ is the number of times topic j has been assigned to some word tokens in document d.

As we do not have a direct sentiment label-document distribution in Reverse-JST, a distribution over sentiment label for document $P(\mathbf{l}|d)$ is calculated as $P(\mathbf{l}|d) = \sum_z P(\mathbf{l}|z, d)P(z|d)$. For all the three models, the probability $P(\mathbf{l}|d)$ will be used to determine document sentiment polarity. We define that a document d is classified as a positive-sentiment document if its probability of positive sentiment label given document $P(\mathbf{l_{pos}}|d)$, is greater than its probability of negative sentiment label given document $P(\mathbf{l_{neg}}|d)$, and vice versa.

4 Experimental Setup

4.1 Dataset Description

Two publicly available datasets, the MR and MDS datasets, were used in our experiments. The MR dataset (also known as the polarity dataset) has become a benchmark for many studies since the work of Pang et al. (2002). The version 2.0 used in our experiment consists of 1000 positive and 1000 negative movie reviews drawn from the IMDB movie archive, with an average of 30 sentences in each document. We also experimented with another dataset, namely *subjective MR*, by removing the sentences that do not bear opinion information from the MR dataset, following the approach of Pang and Lee (2004). The resulting dataset still contains 2000 documents with a total of 334,336 words and 18,013 distinct terms, about half the size of the original MR dataset without performing subjectivity detection.

First used by Blitzer et al. (2007), the MDS dataset contains 4 different types of product reviews taken from Amazon.com including books, DVDs, electronics and kitchen appliances, with 1000 positive and 1000 negative examples for each domain[4].

Preprocessing was performed on both of the datasets. Firstly, punctuation, numbers, non-alphabet characters and stop words were removed. Secondly, standard stemming was performed in order to reduce the vocabulary size and address the issue of data sparseness. Summary statistics of the datasets before and after preprocessing are shown in Table 1.

4.2 Defining Model Priors

In the experiments, two subjectivity lexicons, namely the MPQA[5] and the appraisal lexicon[6], were combined and incorporated as prior information into the model learning. These two lexicons contain lexical words whose polarity orientation have been fully specified. We extracted the words with strong positive and negative orientation and performed stemming in the preprocessing. In addition, words whose polarity changed after stemming were removed automatically, resulting in 1584 positive and 2612 negative words, respectively. It is worth noting that the lexicons used here are fully domain-independent and do not bear any supervised information specifically to the MR, subjMR and MDS datasets. Finally, the prior information was produced by retaining all words in the MPQA and appraisal lexicons that occurred in the experimental datasets. The prior information statistics for each dataset is listed in the last row of Table 1.

In contrast to Lin and He (2009) that only utilized prior information during the initialization of the posterior distributions, we use the prior information in the Gibbs sampling inference step and argue that this is a more appropriate experimental setting. For the Gibbs sampling step of JST and Reverse-JST, if the currently observed word token matches a word in the sentiment lexicon, a corresponding sentiment label will be assigned and only a new topic will be sampled. Otherwise, a new sentiment-topic pair will be sampled for that word token. For LSM, if the current word token matches a word in the sentiment lexicon, a corresponding sentiment label will be assigned and skip the Gibbs sampling procedure. Otherwise, a new sentiment label will be sampled.

[4] We did not perform subjectivity detection on the MDS dataset since its average document length is much shorter than that of the MR dataset, with some documents even having one sentence only.

[5] `http://www.cs.pitt.edu/mpqa/`

[6] `http://lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz`

Table 1: Dataset and sentiment lexicon statistics. (Note:†denotes before preprocessing and * denotes after preprocessing.)

Dataset	# of words					
	MR	subjMR	MDS			
			Book	DVD	Electronic	Kitchen
Corpus size†	1,331,252	812,250	352,020	341,234	221,331	186,122
Corpus size*	627,317	334,336	157,441	153,422	95,441	79,654
Vocabulary†	38,906	34,559	22,028	21,424	10,669	9,525
Vocabulary*	25,166	18,013	14,459	14,806	7,063	6,252
# of lexicon (pos./neg.)*	1248/1877	1150/1667	1000/1352	979/1307	574/552	582/504

Table 2: LSM sentiment classification results.

Aaccuracy (%)

	MR	subjMR	MDS				
			Book	DVD	Electronic	Kitchen	MDS overall
LSM (without prior info.)	61.7	57.9	51.6	53.5	58.4	56.8	55.1
LSM (with prior info.)	74.1	76.1	64.2	66.3	72.5	74.1	69.3
Dasgupta and Ng (2009)	70.9	N/A	69.5	70.8	65.8	69.7	68.9
Li et al.(2009) with 10% doc. label	60	N/A	N/A				62
Li et al.(2009) with 40% doc. label	73.5	N/A					73

5 Experimental Results

5.1 LSM Sentiment Classification Results

In this section, we discuss the sentiment classification results of LSM at document level by incorporating prior information extracted from the MPQA and appraisal lexicon. The symmetry Dirichlet prior β was set to 0.01, and the asymmetric Dirichlet sentiment prior γ was set to 0.01 and 0.9 for the positive and negative sentiment label, respectively. Classification accuracies were averaged over 5 runs for each dataset with 2000 Gibbs sampling iterations.

As can be observed from Table 2, the performance of LSM is only mediocre for all the 6 datasets when no prior information was incorporated. A significant improvement, with an average of more than 13%, is observed after incorporating prior information, especially notable for subjMR and kitchen with 18.2% and 17.3% improvement, respectively. It is also noted that LSM with subjMR dataset achieved 2% improvement over the original MR dataset, implying that the subjMR dataset has better representation of subjective information than the original dataset by filtering out the objective contents. For the MDS dataset, LSM achieved 72.5% and 74.1% accuracy on electronic and kitchen domain respectively, which is much better than the book and DVD domain with only around 65% accuracy. Manually analysing the MDS dataset reveals that the book and DVD reviews often contain a lot of descriptions of book contents or movie plots, which make the reviews from these two domains difficult to classify; whereas in the electronic and kitchen domain, comments on the product are often expressed in a straightforward manner.

When compared to the recently proposed unsupervised approach based on a spectral clustering algorithm (Dasgupta and Ng, 2009), except for the book and DVD domain, LSM achieved better performance in all the other domains with more than 5% overall improvement. Nevertheless, the approach proposed by Dasgupta and Ng (2009) requires users to specify which dimensions (defined by the eigenvectors in spectral clustering) are most closely related to sentiment by inspecting a set of features derived from the reviews for each dimension, and clustering is performed again on the data to derive the final results. In all the Bayesian models studied here, no human judgement is required. Another recently proposed non-negative matrix tri-factorization approach (Li et al., 2009) also employed lexical prior knowledge for semi-supervised sentiment classification. However, when incorporating 10% of labelled documents for training, the non-negative matrix tri-factorization approach performed much worse than LSM, with only around 60% accuracy achieved for all the datasets. Even with 40% labelled documents, it still performs worse than LSM on the MR dataset and slightly outperforms LSM on the MDS dataset. It is worth noting that no labelled documents were used in the LSM results reported here.

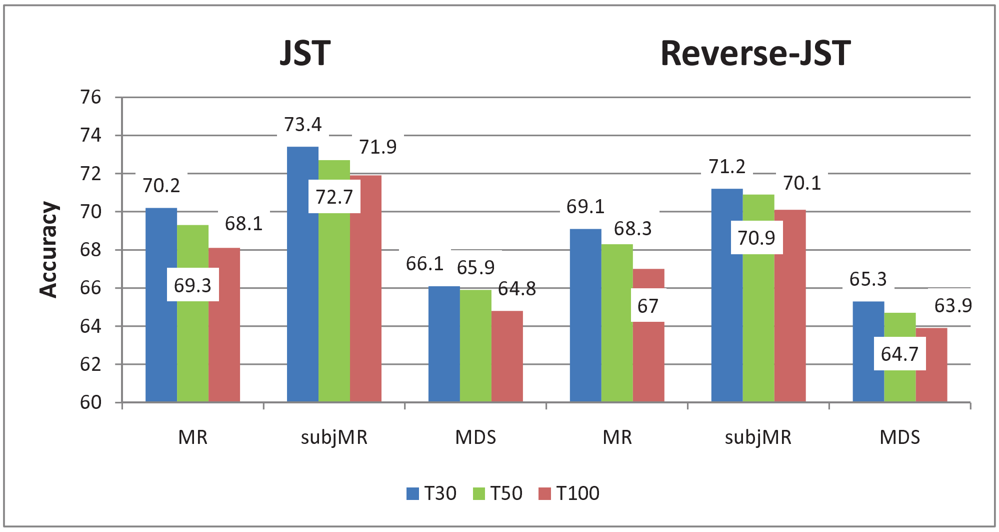


Figure 2: JST and Reverse-JST sentiment classification results with multiple topics.

5.2 JST and Reverse-JST Results with Multiple Topics

As both JST and Reverse-JST model document level sentiment and mixture of topic simultaneously, it is worth to explore how the sentiment classification and topic extraction tasks affect/benifit each other. With this in mind, we conducted a set of experiments on both JST and Reverse-JST, with topic number varying from 30, 50 to 100. The symmetry Dirichlet prior α and β were set to $50/T$ and 0.01 respectively for both models. The asymmetry sentiment prior γ was empirically set to (0.01, 1.8) for JST and (0.01, 0.012) for Reverse-JST, corresponding to positive and negative sentiment prior, respectively. Results were averaged over 5 runs with 2000 Gibbs sampling iterations.

As can be seen from Figure 2 that, for both models, the sentiment classification accuracy based on the subjMR dataset still outperformed the results based on the original MR dataset, where an overall improvement of 3% is observed for JST and about 2% for Reverse-JST. When comparing JST and Reverse-JST, it can be observed that Reverse-JST performed slightly worse than JST for all sets of experiments with about 1% to 2% drop in accuracy. By closely examining the posterior of JST and Reverse-JST (c.f. Equation 4 and 6), we noticed that the count $N_{j,d}$ (number of times topic j associated with some word tokens in document d) in the Reverse-JST posterior would be relatively small due to the factor of large topic number setting. On the contrary, the count $N_{k,d}$ (number of times sentiment label k assigned to some word tokens in document d) in the JST posterior would be relatively large as k is only defined over 3 different sentiment labels. This essentially makes JST less sensitive to the data sparseness problem and the perturbation of hyperparameter setting. In addition, JST encodes an assumption that there is approximately a single sentiment for the entire document, i.e. the documents are usually either mostly positive or mostly negative. This assumption is important as it allows the model to cluster different terms which share similar sentiment. In Reverse-JST, this assumption is not enforced unless only one topic for each sentiment is defined. Therefore, JST appears to be a more appropriate model design for joint sentiment topic detection.

In addition, it is observed that the sentiment classification accuracy of both JST and Reverse-JST drops slightly when the topic number increases from 30 to 100, with the changes of 2% (MR) and 1.5% (subjMR and MDS overall result) being observed for both models. This is likely due to the fact that when the topic number increases, the probability mass attracted under a sentiment-topic pair would become smaller, which essentially creates data sparseness problem. When comparing with LSM, we notice that the difference in sentiment classification accuracy is only marginal by additionally modelling a mixture of topics. But both JST and Reverse-JST are able to extract sentiment-oriented topics apart from document level sentiment detection.

Table 3: Topic examples extracted by JST under different sentiment labels.

Book		DVD		Electronic		Kitchen	
pos.	neg.	pos.	neg.	pos.	neg.	pos.	neg.
recip	war	action	murder	mous	drive	color	fan
food	militari	good	killer	hand	fail	beauti	room
cook	armi	fight	crime	logitech	data	plate	cool
cookbook	soldier	right	cop	comfort	complet	durabl	air
beauti	govern	scene	crime	scroll	manufactur	qualiti	loud
simpl	thing	chase	case	wheel	failur	fiestawar	nois
eat	evid	hit	prison	smooth	lose	blue	live
famili	led	art	detect	feel	backup	finger	annoi
ic	iraq	martial	investig	accur	poorli	white	blow
kitchen	polici	stunt	mysteri	track	error	dinnerwar	vornado
varieti	destruct	chan	commit	touch	storag	bright	bedroom
good	critic	brilliant	thriller	click	gb	purpl	inferior
pictur	inspect	hero	attornei	conveni	flash	scarlet	window
tast	invas	style	suspect	month	disast	dark	vibrat
cream	court	chines	shock	mice	recogn	eleg	power

5.3 Topic Extraction

We also evaluated the effectiveness of topic sentiment captured. In contrast to LDA in which a word is drawn from the topic-word distribution, in JST or Reverse-JST, a word is drawn from the distribution over words conditioned on both topic and sentiment label. As an illustration, Table 3 shows eight topic examples extracted from the MDS dataset by JST, where each topic was drawn from a particular product domain under positive or negative sentiment label.

As can be seen from Table 3, the eight extracted topics are quite informative and coherent, and each of the topics represents a certain product review from the corresponding domain. For example, the positive book topic probably discusses a good cookbook; the positive DVD topic is apparently about a popular action movie by Jackie Chan; the negative electronic topic is likely to be complains regarding data lose due to the flash drive failure, and the negative kitchen topic is probably the dissatisfaction of the high noise level of the *Vornado* brand fan. In terms of topic sentiment, by examining through the topics in the table, it is evident that topics under the positive and negative sentiment label indeed bear positive and negative sentiment respectively. The above analysis reveals the effectiveness of JST in extracting topics and capturing topic sentiment from text.

6 Conclusions and Future Work

In this paper, we studied three closed related Bayesian models for unsupervised sentiment detection, namely LSM, JST and Reverse-JST. As opposing to most of the existing approaches to sentiment classification which favour in supervised learning, these three models detect sentiment in a fully unsupervised manner. While all the three models gives either better or comparable performance compared to the existing approaches on unsupervised sentiment classification on the MR and MDS datasets, JST and Reverse-JST can also model a mixture of topics and the sentiment associated with each topic. Moreover, extensive experiments conducted on the datasets from different domains reveal that JST always outperformed Reverse-JST, suggesting JST being a more appropriate model design for joint sentiment topic detection.

There are several directions we plan to investigate in the future. One is incremental learning of the JST parameters when facing with new data. Another one is semi-supervised learning of the JST model with some supervised information being incorporating into the model parameter estimation procedure such as some known topic knowledge for certain product reviews or the document labels derived automatically from the user-supplied review ratings.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26(3):1–34.

Alina Andreevskaia and Sabine Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of (ACL-HLT)*, pages 290–298.

A. Aue and M. Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 440–447.

S. Dasgupta and V. Ng. 2009. Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 580–589.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 50–57.

Nobuhiro Kaji and Masaru Kitsuregawa. 2006. Automatic construction of polarity-tagged corpus from html documents. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 452–459.

A. Kennedy and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *Proceedings of the Association for Computational Linguistics and the Human Language Technology Conference (ACL-HLT), Short Papers*, pages 257–260.

Tao Li, Yi Zhang, and Vikas Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of (ACL-IJCNLP)*, pages 244–252.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the ACM international conference on Information and knowledge management (CIKM)*.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 432–439.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the conference on World Wide Web (WWW)*, pages 171–180.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, page 271.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Ivan Titov and Ryan McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Aunal Meeting on Association for Computational Linguistics and the Human Language Technology Conference (ACL-HLT)*, pages 308–316.

Ivan Titov and Ryan McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *Proceeding of the International Conference on World Wide Web (WWW 08')*, pages 111–120.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the ACM international conference on Information and Knowledge Management (CIKM)*, pages 625–631.