Exploiting Coherence for the Simultaneous Discovery of Latent Facets and associated Sentiments

Himabindu Lakkaraju* Chiranjib Bhattacharyya† Indrajit Bhattacharya‡ Srujana Merugu*

Abstract

Facet-based sentiment analysis involves discovering the latent facets, sentiments and their associations. Traditional facet-based sentiment analysis algorithms typically perform the various tasks in sequence, and fail to take advantage of the mutual reinforcement of the tasks. Additionally, inferring sentiment levels typically requires domain knowledge or human intervention. In this paper, we propose a series of probabilistic models that jointly discover latent facets and sentiment topics, and also order the sentiment topics with respect to a multi-point scale, in a language and domain independent manner. This is achieved by simultaneously capturing both short-range syntactic structure and long range semantic dependencies between the sentiment and facet words. The models further incorporate *coherence* in reviews, where reviewers dwell on one facet or sentiment level before moving on, for more accurate facet and sentiment discovery. For reviews which are supplemented with ratings, our models automatically order the latent sentiment topics, without requiring seed-words or domain-knowledge. To the best of our knowledge, our work is the first attempt to combine the notions of *syntactic and semantic dependencies* in the domain of review mining. Further, the concept of *facet and sentiment coherence* has not been explored earlier either. Extensive experimental results on real world review data show that the proposed models outperform various state of the art base-lines for facet-based sentiment analysis.

Keywords

Text Mining, Probabilistic Modeling, Review Mining.

1 Introduction

With online expression of sentiment becoming freely available in large volumes, and customers increasingly relying on reviews to decide on products, demand has been growing for opinion mining techniques that help customers in comparing between the various brands, as well as companies to understand market sentiment and improve their products. While early sentiment analysis techniques focused on determining an overall sentiment score for each review, more recent approaches try to discover reasons for satisfaction or dissatisfaction and associate sentiments with specific product features. Since the number of features often runs into hundreds, features are grouped into product facets, and opinion summaries are provided for each facet.

Table 1: Review Snippets for Digital Cameras

The pictures i took during my last trip with this camera were absolutely great. The picture quality is amazing and the pics come out clear and sharp. I am also very impressed with its battery life, unlike other cameras available in the market, the charge lasts long enough. However, I am unhappy with the accessories.
The flash washes out the photos, and the camera takes very long to turn on.

A typical facet-based sentiment analysis algorithm works in two stages, where feature mentions are first identified and grouped into facets, and then sentiment-expressing words and phrases are identified around the facet mentions, and associated with them. Some approaches change the order to use sentiments to identify sentiments. Often, however, this sequential approach can be improved upon.

Consider a first-time prospective camera buyer reading online reviews to decide on a model. Table 1 shows two interesting review snippets. To begin with, she is unaware about many camera features. However, taking cue from common sentiment words such as 'impressed' and 'unhappy' in the first example review, she becomes aware that 'battery life' and 'charge' correspond to an important camera facet, so does 'accessories'. Conversely, she is able to identify new sentiment expressions from camera facets that she is already aware of. For example, she understands that the 'picture quality' facet has been rated positively using the sentiment words 'clear' and 'sharp'. However, she faces a different problem in the second review snippet. Being aware that 'flash' is a camera feature, she realizes that some sentiment is being expressed about it, but she is unsure if 'washing out photos' is desirable, and specifically about

*IBM Research India, {klakkara, srujanamerugu}@in.ibm.com
†Indian Institute of Science, {chiru, indrajit}@csa.iisc.ernet.in

the polarity of the sentiment expressed. However, she observes that it is adjacent to an expression of negative sentiment. Additionally, if overall rating of this review is 2/5, she has enough reason to believe that 'washing out' corresponds to a negative sentiment expressed about the 'flash' facet.

The above example illustrates the different subtasks and dilemmas that are also faced by algorithms for facet based sentiment analysis. First, words denoting facets and sentiments need to be identified, their associations need to be established, and then both classes of words need to be grouped into semantic categories or topics. Additionally, the different sentiment topics need to be ordered. While illustrating the uncertainties involved in these tasks, the example also suggests that they can often be resolved by performing the tasks jointly, rather than sequentially, since they can potentially reinforce each other.

The key to this joint modeling is a combination of syntactic and semantic analysis. The class of facet words, and the class of sentiment words are syntactically distinct, but they are also related to each other through short range syntactical dependencies. For example, in our first review snippet, the words *pictures, picture quality, pics, battery life, accessories* correspond to various facets, while *love, amazing, great, clear, sharp, long, impressed, unhappy* are all accompanying sentiment words. On the other hand, grouping facet words into latent facet topics, and sentiment words into latent sentiment topics, is based on semantic correlations between words.

An important feature of user-generated content such as reviews is *coherence*. When writing a review, users tend to dwell on a particular facet or sentiment level, before moving on to another. In the first review, the user focused on the facet 'picture quality' and a positive sentiment for a few contiguous sentences.

Finally, unlike facets, where it is enough to identify the different facets topics, sentiment topics also need to be ordered. For instance, it is important to identify which groups of words correspond to sentiment ratings *Very Satisfied, Satisfied, Neutral, Dissatisfied* and *Very Dissatisfied*. In our example, it is necessary to identify the sentiment rating associated with 'washing off' for the 'flash' facet. This task typically involves domain knowledge in the form of seed words for individual sentiment ratings, or some other form of human intervention. However, many online merchandizing sites, such as Amazon, Epinions etc., allow users to provide an overall review rating along with their feedback. When such overall ratings are available, the example suggests that it is possible to automatically infer the facet-based sentiment ratings as well.

Motivated by these ideas, in this paper, we develop a sequence of models that jointly identify latent facets and sentiments, their associations, and also the ratings corresponding to the sentiment topics. At the heart of this joint modeling is combination of syntactic and semantic dependencies, similar to the HMM-LDA model [8]. Our model is also able to discover and make use of coherence in reviews, to better identify facet and sentiment topics. Also, by modeling the overall rating, when available, as a response variable depending on the individual sentiments expressed in a review, our model is able to automatically infer the latent sentiment ratings. As a side-effect, this model also allows us to predict overall ratings for new review documents.

Our specific contributions are as follows:

- We propose several models for facet-based sentiment analysis that discover latent facet topics and the corresponding sentiment ratings. All aspects of the model, including the sentiment ratings are learnt in a language-independent manner, without any domain knowledge or expert intervention. To the best of our knowledge, our models are the first to combine the syntactic structure and semantic dependencies leading to fully automated facet-based sentiment analysis.
- We introduce notions of facet and sentiment coherence and model these different types of coherence in customer reviews.
- Using extensive experiments covering all aspects of the task, we demonstrate that our models outperform various state-of-the-art baselines.

The rest of the paper is organized as follows. In Section 2 we discuss the related work. The main contributions are in Section 3 and the experimental evaluation is discussed in Section 5.

2 Related Work

The area of opinion mining has been of great interest since the last decade. Most of the work in this field can be categorized to be attempting to solve one of the following four sub problems. Sentiment Analysis, Feature level opinion mining, Facet(aspect) Extraction and Facet(aspect) based opinion summarization. Though our models discover the latent facets and sentiments, and hence would be able to tackle most of these tasks, our work is mainly targeted at *facet based sentiment analysis*

Facet level sentiment analysis has been of great interest from the past half decade because of its practical utility. This involves extracting the facets and the associated sentiments. Hu and Liu [1] formulated this prob-

lem and applied association mining to extract product features and used a seed set of adjective expanded using wordnet synsets to identify the polarity of the sentiment words, but they do not make any attempts to cluster the product features obtained into appropriate facets. OPINE [2] also tackled the same problem using some rule-based ontologies and relaxation labelling approaches. More recently, Jin and Ho [3] proposed a lexicalized HMM based approach to feature level review mining, their approach integrates linguistic features into automatic learning and is effective in determining the feature terms and opinion words, however since it is based on HMMs training data is needed. However, the main drawback of all these techniques is that they do not cluster the product features into appropriate facets. This is very important in a typical real world scenario where the vocabulary sizes are large.

Mei et. al. [7] proposed Topic Sentiment Model (TSM) which jointly models the mixture of topic and sentiment for weblogs. However this model is based on pLSA [12] and hence has its inherent disadvantages of overfitting and inability to handle unseen data. Though encouraging results are reported on weblogs using this model, it is unclear as to how sentiment towards the discovered topics can be modelled because it is essentially bag-of-words model which does not permit exploiting the co-occurences of topic words with sentiment words. Titov and McDonald further proposed MAS [4] based on MG-LDA [5], a probabilistic model which effectively extracts the ratable facets of a product. However, MAS makes the assumption that atleast one facet is rated in one review which is not a very practical assumption to make. Further, Brody and Elhadad [14] proposed a sentence level topic model to extract facets and identifying the sentiment words using a polarity propagation approach. Though this approach is unsupervised (or requires a weak supervision in the form of seed words), it still treats the task of facet based sentiment analysis as a two stage process. Lin and He [6] proposed a joint sentiment/topic model (JST) which exploits the co-occurrences of topic words with sentiment words in order to achieve better results in terms of analyzing the overall sentiments of reviews. However, since this work does not explicitly model the distinction between the facet words and the sentiment words, it is not directly applicable to facet based sentiment analysis.

3 Generative models for Simultaneous discovery of Facets and Sentiments

In this section, we propose a series of probabilistic models for facet-based opinion summarization. Specifically, the tasks involved are determining which words correspond to the facets and sentiments, then grouping the sentiment and facet words into appropriate topics, and finally ordering the latent sentiment topics so that they correspond to sentiment ratings. This is achieved by introducing three hidden variables for each word in the review document. A class label represents the syntactic category of the word, whether it is a facet word, a sentiment word, or belongs to some other category. Then we have a topic variable for each of the facet and sentiment categories that represents the hidden topic under that category.

In the introduction, we motivated the importance of capturing both short range syntactic structure and long range semantic dependencies. In this section, we interpret them as dependencies among these three hidden variables over the words in the review document, and develop a series of models capturing these dependencies one by one. We start from a basic model that captures both syntax and semantics for identifying latent facet and sentiment topics, to a next one that models coherence in reviews, and then the final model for reviews with overall ratings that is also able to order the sentiment topics.

Table 2: Review Generation Process for FACTS

1. Choose $\theta_d^f \sim Dir(\alpha^f)$
2. Choose $\theta_d^s \sim Dir(\alpha^s)$
3. For each word i
   a. Choose $f_{d,i} \sim Mult(\theta_d^f)$
   b. Choose $s_{d,i} \sim Mult(\theta_d^s)$
   c. Choose $c_{d,i} \sim Mult(\pi^{c_{d,i-1}})$
   d. if $c_{d,i} = 1$, Choose $w_{d,i} \sim Mult(\phi^f_{f_{d,i}})$
      else if $c_{d,i} = 2$, Choose $w_{d,i} \sim Mult(\phi^s_{s_{d,i}})$
      else Choose $w_{d,i} \sim \text{Mult}(\phi^c_{c_{d,i}})$

3.1 Combining Syntax and Semantics: The FACTS Model Our first model, FACTS(FACeT and Sentiment extraction), captures the idea that the syntactic categories of words are dependent through the sentence structure of reviews, while the topic variables have long range semantic dependencies inside a review. This is captured by augmenting the HMM-LDA model [8] for general review mining, where instead of a generic syntactic and semantic class, we are interested in capturing facet, opinion and other 'background' classes. Background classes capture all those words ('trip', 'market', 'I', 'with', 'this' etc) which are not useful in themselves, but are syntactically related to facet and opinion words.

The graphical representation of the FACTS model

is shown in Figure 1(a). A review document d of N_d words, is denoted as $\mathbf{w}_d = \{w_{d,1}, w_{d,2} ... w_{d,N}\}$ where each $w_{d,i}$ is one of the V words in the vocabulary. With each review word $w_{d,i}$, we associate three hidden variables $c_{d,i}$, $f_{d,i}$ and $s_{d,i}$ where each $c_{d,i} \in \{1,2,..C\}$, C being the number of syntactic classes. Words corresponding to $c_{d,i} = 1$ are facet words, those corresponding to $c_{d,i} = 2$ are sentiment expressing words, and other words are background words. Each $f_{d,i}$ takes values from 1 to K^f, each indicating a facet topic, and each $s_{d,i}$ can takes a value from 1 to K^s, indicating a particular sentiment topic. Each facet and sentiment topic is associated with a distribution over words in the vocabulary, ϕ_t^f for the t^{th} facet topic and ϕ_k^s for the k^{th} sentiment topic, respectively. Each class other than facet and sentiment classes have their own distribution over words ϕ_j^c, $j \neq 1, 2$. The facet and sentiment classes have indirect word distributions through their respective topics. The complete generative process for each review is described in Table 2.

Further, we assume that the various multinomial distributions are generated using symmetric Dirichlet priors. The document-specific distributions, θ_d^f over facet topics, and θ_d^s over sentiment topics, are drawn from Dir(α^f) and Dir(α^s) respectively. Similarly, the topic distributions, ϕ^f for facets and ϕ^s for sentiments, are drawn from Dir(β^f) and Dir(β^s) respectively. The rows of the transition matrix for the HMM, π, are drawn from Dir(γ), the class distributions ϕ^c are drawn from Dir(δ).

3.2 Addressing Coherence: The CFACTS model The FACTS model captures the syntactic dependencies between the facet, sentiment and other background classes. However, the only dependence among the latent topic variables are long-range, and they are not influenced by other local topics. In contrast, as motivated in the introduction, reviews are often characterized by *Facet Coherence*, where users comment about a particular facet in contiguous text fragments, and *Sentiment Coherence*, where sentiments expressed in contiguous text fragments are related.

Our next model, CFACTS (Coherence based FACTS), captures this idea by extending the FACTS model by introducing dependencies between the sentiment and facet topic variables for neighboring words. For this model, we first introduce a *window*, which is a contiguous sequence of words, as the basic unit of coherence. All facet words within a window are assumed to be derived from the same facet topic $f_{d,x}$, and all sentiment words from the same sentiment topic $s_{d,x}$. Depending upon the corpus, this may be a sentence, or a group of words or phrases split by delimiters. The sentiments and facets for adjacent windows may still be dependent for a specific review document in multiple ways. This is captured by introducing a multinomial variable $\psi_{d,x}$ for each window x of a review d.

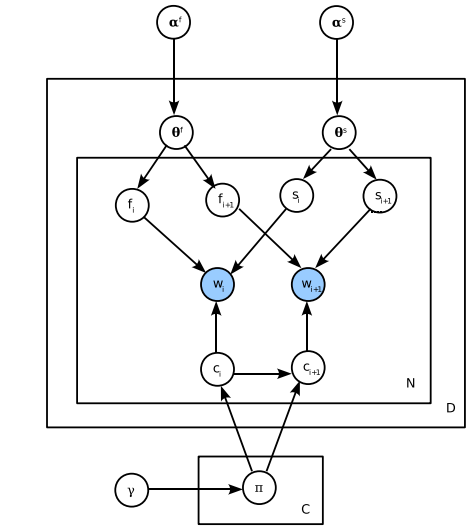
- $\psi_{d,x} = 0$ indicates that both the facet and sentiment topics of the window x are the same as those of the previous window $x-1$: $f_{d,x} = f_{d,x-1}$ and $s_{d,x} = s_{d,x-1}$
- $\psi_{d,x} = 1$ indicates the sentiment topic of window x is the same as that of the previous window, but the facet topic is independent : $f_{d,x} \sim \theta_d^f$ and $s_{d,x} = s_{d,x-1}$
- $\psi_{d,x} = 2$ both the facet and sentiment topics of window x are independent of the previous window : $f_{d,x} \sim \theta_d^f$ and $s_{d,x} \sim \theta_d^s$

Table 3: Review Generation Process for CFACTS

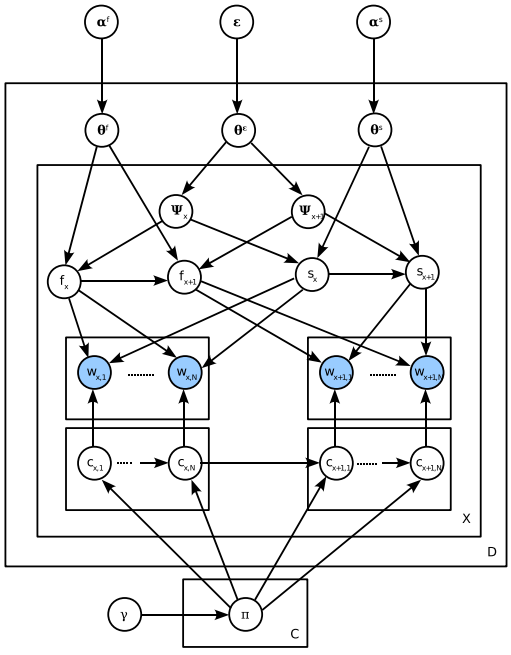
1. Choose $\theta_d^f \sim Dir(\alpha^f)$
2. Choose $\theta_d^s \sim Dir(\alpha^s)$
3. Choose $\theta_d^\epsilon \sim Dir(\epsilon)$
4. For each window $x \in \{1...X_d\}$
   a. Choose $\psi_{d,x} \sim Mult(\theta_d^\epsilon)$
   b. if ($\psi_{d,x}$=0)
      Choose $f_{d,x} = f_{d,x-1}$ and $s_{d,x} = s_{d,x-1}$
      else if ($\psi_{d,x}$=1)
      Choose $f_{d,x} \sim \theta_d^f$ and $s_{d,x} = s_{d,x-1}$
      else Choose $f_{d,x} \sim Mult(\theta_d^f)$ and $s_{d,x} \sim Mult(\theta_d^s)$
   c. For each word i in the window x
      i. Choose $c_{d,x,i} \sim Mult(\pi^{c_{d,x,i-1}})$
      ii. if $c_{d,x,i} = 1$, Choose $w_{d,x,i} \sim Mult(\phi^f_{f_{d,x}})$
          else if $c_{d,x,i} = 2$, Choose $w_{d,x,i} \sim Mult(\phi^s_{s_{d,x}})$
          else Choose $w_{d,x,i} \sim Mult(\phi^c_{c_{d,x,i}})$

Notice that we intentionally avoided the unlikely scenario where users express contradicting opinions about the same facet in a single review. The specific values of $\psi_{d,x}$, sampled from a multinomial distribution $Mult(\theta_d^\epsilon)$ determines the specific scenario for any window x, where θ_d^ϵ is sampled once for each review. Observe that the earlier FACTS model comes out as a special case of CFACTS with no coherence, which is captured by a window length of 1 word, and $\psi_{d,x} = 2$.

The CFACTS generative process for each review document is given in the Table 3 and Figure 1(b) presents the corresponding graphical model. Notice that now the facet topic $f_{d,x}$ and the sentiment topic $s_{d,x}$ are sampled once and reused for all words within a window x.



(a)



(b)

Figure 1: a. FACTS model b. CFACTS model

3.3 Inferring facet-level sentiment ratings: CFACTS-R The FACTS and CFACTS models generate review documents with different types of syntactic and semantic dependencies between the syntactic class variables and the facet and sentiment topic variables. However, the topic variables are categorical, and any two topics are equally different for both facets and sentiments. While this is enough for the facet topics, the sentiment topics are naturally ordered — A rating of 1 is closer to rating of 2 than to 3. The task of ordering the sentiment topics, has to be undertaken outside of the model, either using domain specific seed-words for individual levels, or some other domain knowledge. This task can become unmanageable as the finer levels of sentiments are required.

Table 4: Review Generation Process for CFACTS-R

% Follows steps 1-4 as for CFACTS
5. $r_d \sim Normal(\eta^T \bar{s_d}, \sigma^2)$

Clearly, as motivated in our example, the overall review score is dependent on the individual sentiment topics in any review document. We can model the overall score as a response variable with a probability distribution determined by latent sentiment topics of all relevant words in the review. Accordingly, our next model, which we call CFACTS-R (CFACTS with Rating), extends the CFACTS review-document generation process by an additional step. Once all the words and their facet or sentiment topics have been generated over all windows for a review d exactly as in Table 3, the response variable r_d for the review is generated using a normal linear model as shown in Table 4, where $\bar{s_d}$ is a vector of length K^s and an element i of this vector corresponds to the empirical probability of the i^{th} sentiment topic in the review document d i.e the i^{th} element of this vector is given by $\bar{s_{di}} := (1/X_d) \sum_{x=1}^{X_d} I(s_{d,x} = i)$, I(.) being the indicator function and the vector η comprises of the coefficients of each of these empirical probabilities which will be estimated. This kind of normal linear model has been used alongside LDA in [13].

Modeling the overall rating this way leads to two different benefits. Firstly, when overall review ratings are observed, the latent sentiment ratings can be automatically inferred. On the other hand, once the regression coefficients η have been learned from a review collection, the overall rating can now be inferred for new reviews based on the sentiment levels of words in the review. We evaluate both these aspects in our experiments in Section 5.

The resultant CFACTS-R model captures all the requirements that we had motivated for the joint model. It appropriately combines syntax and semantics to jointly identify facet and sentiment words in a review, and also their latent topics. It models coherence in reviews to capture local continuity of facet and sentiment topics. Finally, it models the dependence of the overall review rating based on the individual sentiment ratings. Of course, it is also possible to

imagine variants that model the overall review rating, but do not consider coherence. We consider such a model, which we call FACTS-R, in our experiments.

4 Inference Using Gibbs sampling

In this section, we present the inference algorithms for the proposed models. The inference task is to compute the conditional distribution over the set of hidden variables for all words in the collection of review documents. Exactly computing this distribution is intractable. Here we perform approximate inference using collapsed Gibbs Sampling [8], where the conditional distribution is computed for each hidden variable based on the current assignment of all the other hidden variables, and integrating out the other parameters in the model. The inference algorithm then repeatedly samples values for the hidden class, facet topic and sentiment topic for each word in the collection from this conditional distribution until convergence. Due to space limitations, we only describe the derived conditional distributions for the different hidden variables.

In the following expressions, $\boldsymbol{W}$ denotes all the words present in the corpus, $\boldsymbol{F}$ denotes all the facet topic assignments, $\boldsymbol{S}$ refers to all the sentiment topic assignments, $\boldsymbol{C}$ to all the class assignments, and $\boldsymbol{R}$ denote the set of overall ratings of all the documents. Any of these variables subscripted with $\boldsymbol{-(d,i)}$ indicates that the i^{th} word of the d^{th} document is excluded. Similarly, subscript $\boldsymbol{-(d,x)}$ indicates that the x^{th} window of the d^th document is excluded.

We first present the conditional distributions for the most general model, which is CFACTS-R, and then discuss the differences for CFACTS and FACTS.

Inference for CFACTS-R Recall that in the coherent models, CFACTS and CFACTS-R, a document is modeled as a collection of windows, each window being a collection of words. All words within a window have the same facet and sentiment topics, $f_{d,x}$ and $s_{d,x}$. Also these are dependent through the $\psi_{d,x}$ variable. Therefore, in the sampling algorithm, for each window in each document, we sample (f, s, ψ) as a block, along with sampling the class for each word individually. To keep the notations compact, we further use $\boldsymbol{H}$ to denote ($\boldsymbol{F}$,$\boldsymbol{S}$,$\boldsymbol{\psi}$) Then the conditional distribution for $H_{(d,x)}$ looks as follows:

$$P(H_{(d,x)} = (t,q,l)|\boldsymbol{H_{-(d,x)}}, \boldsymbol{C}, \boldsymbol{W}, \boldsymbol{R})$$

$$\propto P(\psi_{(d,x)} = l|\boldsymbol{\psi_{-(d,x)}})\times$$

$$\underbrace{P(f_{(d,x)} = t, s_{(d,x)} = q|\psi_{(d,x)}, \boldsymbol{H_{-(d,x)}})}_{g_1}\times$$

$$P(\boldsymbol{w}_{(d,x)}|\boldsymbol{H}, \boldsymbol{C}, \boldsymbol{W_{-(d,x)}})\times$$

$$P(\boldsymbol{r}_d|\boldsymbol{H}, \boldsymbol{C}, \boldsymbol{W_{-(d,x)}})$$

$$\propto \left(n_{d,(\cdot)}^{l,-(d,x)} + \epsilon\right) \times g_1 \times \prod_{\substack{i=1 \\ c_{d,x,i}=1}}^{N_{d,x}} \frac{\left(n_{(\cdot),v}^{t,-(d,x,i)} + \beta^f\right)}{\left(\sum_{r=1}^{V}(n_{(\cdot),r}^{t,-(d,x,i)} + \beta^f)\right)}$$

$$\times \prod_{\substack{i=1 \\ c_{d,x,i}=2}}^{N_{d,x}} \frac{\left(n_{(\cdot),v}^{q,-(d,x,i)} + \beta^s\right)}{\left(\sum_{r=1}^{V}(n_{(\cdot),r}^{q,-(d,x,i)} + \beta^s)\right)} \times p_d^r$$

where g_1 is computed as follows for various choices of l.

	g_1
$l = 0$	1 if $t = f_{d,x-1}$ and $q = s_{d,x-1}$, 0 otherwise
$l = 1$	$\left(n_{d,(\cdot)}^{t,-(d,x)} + \alpha^f\right)$ if $q = s_{d,x-1}$, 0 otherwise
$l = 2$	$\left(n_{d,(\cdot)}^{t,-(d,x)} + \alpha^f\right) \times \left(n_{d,(\cdot)}^{q,-(d,x)} + \alpha^s\right)$

The word $w_{d,x,i}$ corresponds to the v^{th} word of the vocabulary. $n_{d,(\cdot)}^{l,-(d,x)}$ is the number of windows in the document d for which $\psi_{d,x}$ takes the value l. $n_{d,(\cdot)}^{t,-(d,x)}$ is the number of windows in the document d assigned to the facet topic t. $n_{d,(\cdot)}^{q,-(d,x)}$ is the number of windows in the document d assigned to the sentiment topic q. $n_{(\cdot),v}^{t,-(d,x,i)}$ is the number of times v^{th} word in the vocabulary is assigned topic t, and

$$p_d^r = P(\boldsymbol{r}_d|\boldsymbol{H}, \boldsymbol{C}, \boldsymbol{W_{-(d,x)}}) \propto e^{-(r_d-\mu_d)^2/2\sigma^2}$$

$$\text{with } \mu_d = \eta_q \frac{n_{d,(\cdot)}^{q,-(d,x)} + 1}{X_d} + \sum_{y \neq q} \eta_y \frac{n_{d,(\cdot)}^{y,-(d,x)}}{X_d}$$

(Note that all these counts which are superscripted with -(d,x) or -(d,x,i) do not include the corresponding instances).

The other required conditional distribution is for the class variable for the i^{th} word of the x^{th} window belonging to the d^{th} document. This distribution takes the following form:

$$P(c_{(d,x,i)} = u|\boldsymbol{C_{-(d,x,i)}}, \boldsymbol{H}, \boldsymbol{W})$$

$$\propto \underbrace{P(w_{(d,x,i)}|\boldsymbol{F}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W_{-(d,x,i)}})}_{g_2} P(c_{(d,x,i)}|\boldsymbol{C_{-(d,x,i)}})$$

where g_2 is computed as follows for various choices of u.

	g_2
$u = 1$	$\frac{\left(n_{(\cdot),v}^{f_{d,x,i},-(d,x,i)} + \beta^f\right)}{\left(\sum_{r=1}^{V}(n_{(\cdot),r}^{f_{d,x,i},-(d,x,i)} + \beta^f)\right)}$
$u = 2$	$\frac{\left(n_{(\cdot),v}^{s_{d,x,i},-(d,x,i)} + \beta^s\right)}{\left(\sum_{r=1}^{V}(n_{(\cdot),r}^{s_{d,x,i},-(d,x,i)} + \beta^s)\right)}$
otherwise	$\frac{\left(n_{(\cdot),v}^{u,-(d,x,i)} + \delta\right)}{\left(\sum_{r=1}^{V}(n_{(\cdot),r}^{u,-(d,x,i)} + \delta)\right)}$

Also, $P(c_{(d,x,i)} = u|\boldsymbol{C_{-(d,x,i)}}) =$

$$\frac{(n_u^{c_{d,x,i-1}} + \gamma)}{n_\cdot^u + I(c_{d,x,i-1} = u) + C\gamma} \times$$

$$(n^u_{c_{d,x,i+1}} + I(c_{d,x,i-1} = u)I(c_{d,x,i+1} = u) + \gamma)$$

$n_u^{c_{d,x,i-1}}$ is the number of transitions from class $c_{d,x,i-1}$ to the class u, and all counts of transitions exclude transitions both to and from class u. I$(\cdot)$ is an indicator function, taking the value 1 when its argument is true, and 0 otherwise.

Inference for CFACTS The CFACTS model is almost identical to the CFACTS-R model, the only difference being that CFACTS does not consider the overall review ratings $\mathbf{R}$. Accordingly, the only difference between the conditional distribution of $H_{(d,x)}$ for CFACTS is that it does not depend on $\mathbf{R}$, and the distribution looks almost identical to that for CFACTS-R with the $p_d^r = P(\boldsymbol{r}_d|\boldsymbol{H}, \boldsymbol{C}, \boldsymbol{W_{-(d,x)}})$ terms missing.

Inference for FACTS Recall that FACTS model can be obtained as a special case of the CFACTS model, where all the $\psi_{d,x}$ values are 2, and the window size is 1 word. In other words, the FACTS conditional distributions can be derived from the CFACTS update rules in a straight-forward manner, by setting each word to be one window. Also, the implication of $\psi_{d,x} = 2$ is that the facet topic $f_{d,i}$ and the sentiment topic $s_{d,i}$ for word i in document d become independent of each other. Accordingly, for FACTS, we sample $P(f_{(d,i)} = t|\boldsymbol{F_{-(d,i)}}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W}))$ and $P(s_{(d,i)} = q|\boldsymbol{S_{-(d,i)}}, \boldsymbol{F}, \boldsymbol{C}, \boldsymbol{W})$ independently. The distribution for $f_{(d,i)}$ looks as follows:

$$P(f_{(d,i)} = t|\boldsymbol{F_{-(d,i)}}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W})$$

$$\propto \left(n_{d,(\cdot)}^{t,-(d,i)} + \alpha^f\right) \frac{\left(n_{(\cdot),v}^{t,-(d,i)} + \beta^f\right)}{\left(\sum_{r=1}^{V}(n_{(\cdot),r}^{t,-(d,i)} + \beta^f)\right)} \text{ if } c_{d,i} = 1$$

$$\text{or } \left(n_{d,(\cdot)}^{t,-(d,i)} + \alpha^f\right) \text{ if } c_{d,i} \neq 1$$

where the counts are similar to CFACTS-R, except that they are only over documents and words, and do not consider windows. The counts for $f_{(d,i)}$ include only those words for which class is 1. The conditional distribution for $s_{(d,i)}$ looks very similar to this, with the difference being that all the counts are now taken only with respect to class 2. The conditional distribution for the class distribution again takes a form very similar to that for FACTS and CFACTS-R, with the difference that counts are only taken only over documents and words.

5 Experimental Evaluation

In this section, we report experimental evaluation of the proposed models for the task of facet-based sentiment analysis on real world review datasets. In addition to the FACTS, CFACTS and CFACTS-R models, we also consider the FACTS-R model, which automatically infers sentiment ratings using the review score, but does not capture coherence. The overall task involves two important sub-tasks, discovering the latent facets and the latent sentiment ratings. We evaluate our models for these two sub-tasks as well. All models for facet discovery that we are aware of, perform only a qualitative analysis. We evaluate the discovered facets qualitatively, and in addition provide a thorough quantitative evaluation. In addition to facet-level sentiment analysis, by virtue of modeling sentiment distributions at a review level, the models are able to perform review-level sentiment analysis as well. We evaluate the models also for this task. It is very important to note that no single baseline model performs this wide array of tasks across the review mining spectrum. Accordingly, we compare against different state-of-the-art baselines for these various tasks.

Datasets: We experiment with the product review data crawled from Amazon in November 2009. We crawled reviews under the following product categories: Digital Cameras (61,482 reviews), Laptops (10,011 reviews), Mobile Phones (6,348 reviews), LCD TVs (2,346 reviews) and Printers (2,397 reviews) [1]. The crawled reviews were preprocessed to remove html tags, and segmented into sentences. Note that we did not apply stemming or stop-word removal to preserve syntactic structure of sentences.

Default Parameter Settings: For all proposed models, we ran the collapsed gibbs sampling algorithms for 1000 iterations. The hyper-parameters were initialized as: $\alpha^f = 50/K^f$, $\alpha^s = 50/K^s$, $\beta^f = \beta^s = \gamma = \delta = 0.1$. The default window size for the CFACTS and CFACTS-R models is 1 sentence. We present evaluations over varying window sizes at the end of this section.

5.1 Facet Discovery

In this section, we evaluate our models for the task of facet discovery by considering the quality of the extracted facets. We perform evaluations on all five product categories of the Amazon dataset.

Baseline As our baseline, we consider the LDA model [9], which is the state-of-the-art model for unsupervised topic discovery.

Evaluation We analyze the facets generated by different models qualitatively as well as quantitatively.

Qualitative Evaluation: The facets extracted by the

[1] http://mllab.csa.iisc.ernet.in/downloads/reviewmining.html

Table 5: Top words from Facet Topics for Digital Camera Review Corpus

Model	Topic Label	Top Words	Topic Label	Top Words
CFACTS-R	Price	fit, purse, pocket, pay, worth	Display	digital, viewfinder, shots, lens, clarity
(all topics)	Ease of use	ease, weight, casing, digicam, travel	Battery life	aa, batteries, life, ease, charge
	Picture quality	shots, video, camera, images, pics	Portability	travel, ease, bags, portability, straps
	Accessories	charger, cable, battery, controls, button	Features	lens, memory, point-and-shoot, software
CFACTS	Battery life	battery, charge, shutter, aa batteries, alkaline	Ease of use	ease, use, design, color, grip
(all	Accessories	charger, camera, cable, tripod, shutter button	Display	slr, lcd, viewfinder, display, point-and-shoot
topics)	Picture quality	images, clarity, camera, brightness, focus	Price	price, worth, discount, warranty, fit
	Features	zoom, nikon, face recognition, redeye, memory	Portability	ease, portability, size, lightweight, travel
FACTS-R	Accessories	buttons, tripod, controls, batteries, purse	-	memory, quality, purchase, warranty, cams
(5 out of	Lens	shutter, minolta, camera, point-and-shoot	Picture quality	pictures, quality, images, resolution, sharp
8 topics)	Portability	range, size, weight, bag, design		
FACTS	Lens	shutter, lens, camera, point-and-shoot	Picture quality	pictures, quality, images, resolution, sharp
(5 out of	Portability	range, size, weight, bag, design	Accessories	buttons, controls, charger, tripod, purse
8 topics)	-	pics, shots, range, ease, straps		
LDA	Accessories	replace, charger, reader, digicam, easy	Picture quality	images, camera, pics, like, good
(4 out of	-	take, shoot, carry, great, easy	-	charger, lens, awful, camera, shutter
9 topics)				

various models in case of digital camera reviews are recorded in Table 5. The top words column of the table highlights the top 5 words for each facet. Depending upon the words encountered in a particular topic, each topic is assigned a topic label (column 2 of Table 5) manually to bring out the notion of the facet being captured by that particular topic. It can also be seen that some of the topic label fields are left unlabeled indicating that these topics were not coherent enough to correspond to any particular facet.

Quantitative Evaluation: In order to quantitatively evaluate the quality of the facets extracted, we make use of the *structured ratings*[1] available on Amazon. *Structured ratings* is a feature of amazon which allows users to list out the facets of interest for each product and allow them to rate these facets explicitly. We compare the facets extracted by our models with the facets explicitly listed by users on the structured ratings of amazon. For instance, in the case of digital cameras, the facets listed on amazon are {Battery life, Display, Ease of use, Picture quality, Portability, Features}. We propose the following two metrics to assess the quality of the facets generated by our models.

- *Facet Coverage* measures the fraction of extracted facets that actually correspond to those listed on Amazon structured ratings.

- *Facet Purity* for a specific facet measures the fraction of the top words (words constituting about 70% of the probability mass of the word distribution for the facet) in the facet that actually correspond to the labeled product attribute.

[1] http://www.amazon.com/gp/structured-ratings/product/{product-id}

Table 6: Experimental Results for evaluating the Quality of Facet Extraction

Corpus	Model	Facet Coverage(%)	Topic Purity(%)
Digital Cameras	CFACTS-R	**100**	80.18
	CFACTS	**100**	**84**
	FACTS-R	33	74.73
	FACTS	33	72.28
	LDA	16.67	44.37
Laptops	CFACTS-R	**83.33**	**87.09**
	CFACTS	**83.33**	**87.09**
	FACTS-R	33.33	74.19
	FACTS	33.33	77.41
	LDA	33.33	45.16
Mobile Phones	CFACTS-R	**80**	**91.48**
	CFACTS	**80**	89.36
	FACTS-R	40	74.46
	FACTS	40	80.85
	LDA	40	40.42
LCD TVs	CFACTS-R	**80**	78.94
	CFACTS	**80**	**84.21**
	FACTS-R	60	68.42
	FACTS	60	65.78
	LDA	40	36.84
Printers	CFACTS-R	**100**	79.31
	CFACTS	**100**	**84.48**
	FACTS-R	75	75.86
	FACTS	75	72.41
	LDA	75	36.76

Discussion The results of the qualitative and quantitative evaluation are highlighted in the Tables 5 and 6 respectively. Because of space constraints, we focus only on the qualitative evaluation of the digital camera reviews in Table 5. For all these experiments on the facet-discovery, the number of opinion topics (K^s) is set to 2 in case of all the models. It can be seen that CFACTS and CFACTS-R outperform the other

Table 7: Seed word list for opinion words

Polarity	Seed Words
Highly Positive	great, amazing, awesome, excellent, brilliant
Positive	good, easy, happy, love, like
Neutral	fine, enough, ok, okay
Negative	bad, poor, hate, slow, clumsy
Highly Negative	pathetic, terrible, awful, disappointing, worst

Table 8: Accuracy of the word and sentiment polarity

Model	Word Acc (%)	Sentence Acc (%)
CFACTS-R	77.68	80.54
CFACTS	**78.22**	**81.28**
FACTS-R	72.02	72.25
FACTS	72.98	75.72
JST	73.14	76.18

models, highlighting the importance of coherence. The facet coverage goes to as high as 100% for digital camera corpus for CFACTS and CFACTS-R. Further, all of the models significantly outperform the baseline LDA model, showing the importance of considering syntax for the review mining task.

5.2 Sentiment Analysis We next evaluate the models for the task of sentiment analysis. Recall that our proposed models associate sentiment with different units of a review document. The FACTS and the FACTS-R models assign sentiment topics to each word belonging to the sentiment class, while the coherent models CFACTS and CFACTS-R associate a single sentiment with windows (or sentences) within a review. We evaluate sentiment discovery both at a word and sentence level.

Baseline As a baseline, we consider the joint sentiment topic model for sentiment analysis (JST) [6], which reports state of the art accuracy for sentiment analysis tasks. We use our own gibbs sampling based implementation of JST. JST is a two-level model for topics and sentiments. But it does not capture the syntactic notions of these entities, and therefore tags all words in a review (even the non-opinion words) with some sentiment label. Considering this aspect, for a fair comparison, we restricted this task of determining the polarity of the words only to adjectives and verbs.

While the FACTS-R and CFACTS-R models are automatically able to assign levels to sentiment topics by considering the review score, the remaining models, FACTS and CFACTS and also the baseline JST, are only able to identify different sentiment groups and not their levels. Therefore, these models are additionally provided with seed words for each of the sentiment levels. Table 7 records these seed words. For JST model, we set the number of opinion topics to 3, (positive, negative and neutral) and additionally seeded the neutral topic with a set of 10 neutral verbs and adjectives such as {okay, fine, enough, ok, bought, using, went}. For the proposed model, each of the sentiment words are assigned to any of 2 sentiment levels (positive and negative), while all non-opinion words are tagged as neutral. We next describe sentiment evaluation results at word and sentence levels.

5.2.1 Word-level Sentiment Evaluation We benchmark the performance of our models against the sentiment levels obtained from SentiWordNet [2]. We use the scores associated with each word by *SentiWordNet*, and tag each word with the level corresponding to the highest score. For instance, for the word 'bad', with scores assigned by SentiWordNet being {Positive: 0 Neutral: 0.375 Negative: 0.625 }, we tag it as negative.

Discussion The results are shown in the second column of Table 8. As can be seen, CFACTS and CFACTS-R models outperform the other models and the baseline. Also, the CFACTS-R model was not provided with any seeding, but still its performance is on par with the CFACTS model. This demonstrates that usefulness of ratings for automatically inferring sentiment levels. Further, CFACTS and CFACTS-R models demonstrate that the coherence is indeed very useful in discovering sentiments. Further analysis of the results revealed that the accuracies of FACTS and FACTS-R are lower because they could not capture the positive/negative sentiments correctly all the time. But by modeling syntactic classes, neutrality is captured very well by these models. Their overall performance is similar to JST. On the other hand, JST does better in distinguishing between positive and negative polarity but is unable to capture neutrality well.

5.2.2 Sentence-level Sentiment Evaluation Since no gold-standard is available for sentiment levels for sentences, we manually created a labeled dataset for evaluation. We considered two levels of sentiment, positive and negative. From the complete review dataset, we extracted a subset of about 8,012 sentences and manually labeled them as positive or negative. Recall that CFACTS and CFACTS-R directly provide sentiment levels for sentences. For the other models, FACTS, FACTS-R and JST, we associate each sentence with the polarity associated with majority of the sentiment words. For all models we set the number of opinion topics to 2.

Discussion The results are presented in the third column of Table 8. Again, CFACTS and CFACTS-R outperform the rest of the models. It is evident in here too, that the concept of facet and sentiment level coherence

[2] http://sentiwordnet.isti.cnr.it/

can capture the polarities better. CFACTS performs slightly marginally better than CFACTS-R, using the benefit of the seed words, suggesting that seed words contain more sentiment information compared to the overall review rating. However, providing seed words often involve significant effort and domain knowledge, which can be avoided for CFACTS-R, with minimal loss in accuracy. The significance of this is likely to grow as finer granularities of sentiments are desired.

5.3 Facet based Sentiment Analysis In this section, we focus on the overall task of facet-based sentiment analysis, that involves facet and sentiment rating discovery as sub-tasks. More specifically, we evaluate the accuracy of the (facet,sentiment) pairs extracted by our proposed models. We evaluate this accuracy both at the level of sentences and reviews.

To create a gold-standard for evaluation, we took a set of 1500 reviews from the complete Amazon dataset, and manually labeled each of these reviews with (facet,opinion) pairs. For evaluating the facet and sentiment discovery at sentence level, we further labeled a subset of these 1500 reviews with (facet,opinion) pairs for each sentence. To evaluate each model, we measure the accuracy of the extracted (facet,sentiment) pairs with respect to the gold-standard.

We compare the proposed models, FACTS, CFACTS, FACTS-R and CFACTS-R against two different baselines, one that is based on frequent item-set mining, and the other based on LDA. We have seen that apart from CFACTS and CFACTS-R, the other models do not generate facets that correspond well to the actual product attributes. This is true for the baselines as well. To get around this, we provide 2 seed-words per facet topic for a subset of the actual product attributes to establish this correspondence. Note that this seeding is provided only for FACTS, FACTS-R and the two baselines, and is not required for CFACTS and CFACTS-R. A similar seeding is required for sentiment topics in case of FACTS, CFACTS and the two baselines. However, FACTS-R and CFACTS-R can do without needing this, by considering the overall review ratings.

Baselines Before moving to the results, we briefly describe the two baselines that we use.

Frequent itemset based Facet/Sentiment Miner (FIFS): This algorithm, proposed for feature extraction and opinion identification [1], is a state-of-the-art method for associating features and sentiments. It extracts the most frequent noun words appearing in the review corpus and designates them as features. Separately, all adjectives are treated to be opinion words. Then, to detect the polarity of these opinion words, the algorithm starts with some positive and negative polarity seed words, and uses wordnet [10] synsets to determine the polarity of the other opinion words. We use our own implementation of this algorithm. But, this algorithm does not perform *grouping* of the related feature terms into facets. To do this, we post-process the output using seed-words for a subset of facets, and a PMI (Point-wise mutual information) based metric. Then for each feature term e extracted by the algorithm, we compute the similarity with each of these seed sets S_n: $sim(S_n, e) = \sum_{s \in S_n} log_2 \frac{P(s,e)}{P(s)P(e)}$, where $P(s, e)$ is the probability that s and e occur together within a span of two sentences. If $sim(S_n, e) > \epsilon$ (where ϵ signifies the threshold), we add it to the corresponding facet entity.

LDA based Facet/Sentiment Miner (LFS) - As our second baseline, we use an LDA [9] based technique which discovers both the facets and sentiment topics. We tag the words in the reviews with their POS using the stanford POS tagger [11]. We restrict noun words to belong to certain topics which we call the facet topics, and adjectives to belong to positive and negative sentiment topics and all the other words to belong to background topics.

Discussion The results of this experiment are shown in Table 9. The numbers are averaged over all the different product types — the results over the corpora for different products did not vary much. The coherent models, CFACTS and CFACTS-R, significantly outperform all the other models both in facet identification and sentence level polarity detection. This demonstrates the benefit of modeling *coherence* for review mining. FACTS and FACTS-R do moderately well in discovering facets, but falter in the task of polarity detection when reviews have both positive and negative opinion words (as users may like certain facets and dislike others), This shows that capturing document level co-occurrence is not enough to determine the sentiment levels for words. However, all the four proposed models outperform the two baseline algorithms. The LDA based miner can capture facets reasonably correctly, but does not perform well for identifying sentiment levels.

Finally, it needs to be stressed that although CFACTS performs somewhat better than CFACTS-R, this is only by virtue of using seeding for both facet and sentiment topics. CFACTS-R is the only model that does not require any seeding for identifying either facets that correspond to actual product features, or different levels of sentiments. This makes it ideally suited for mining large review corpora with many product attributes at fine levels of sentiment.

5.4 Document Level Sentiment Analysis Though the primary goal of the proposed models is to perform facet-level sentiment analysis, they can also be

Table 9: Experimental Results for Facet Extraction and Sentiment Detection at Sentence level

Model	Two sentiment topics				Five sentiment topics			
	Facet		Polarity	(facet,sentiment)	Facet Identification		Polarity	(facet,sentiment)
	Precision(%)	Recall(%)	Acc(%)	Identification	Precision(%)	Recall(%)	Acc(%)	Identification
CFACTS-R	**83.42**	80.08	84.10	79.11	81.92	80.00	78.41	77.19
CFACTS	83.33	**81.12**	**84.21**	**82.18**	**82.52**	**80.72**	**79.73**	**78.02**
FACTS-R	79.80	80.01	71.18	70.84	79.26	80.18	67.78	65.11
FACTS	79.97	80.28	72.23	71.92	79.81	80.32	67.71	66.32
LFS	70.12	71.90	69.80	69.34	70.12	71.78	65.98	65.72
FIFS	71.22	67.81	64.72	62.30	71.22	67.81	61.88	61.01

used for analyzing sentiments for reviews as a whole. Here, we briefly demonstrate their use for two different tasks -

5.4.1 Determining the polarity of review documents We evaluate the performance of our models on the following two tasks - *binary classification* - classifying a review as positive or negative and *five-class classification* - classifying a review into one of the five sentiment levels - {Highly positive(5), Positive(4), Neutral(3), Negative(2), Highly Negative(1)}.

Baseline As a baseline, we consider the joint sentiment topic model for sentiment analysis (JST)[6], which reports state of the art accuracy for this task. We use our own gibbs sampling based implementation of JST.

Experiments Each of our models estimate θ_d^s which is a document level sentiment topic distribution. Further, the baseline JST also determines this kind of document-sentiment distribution. So, we label each document with a sentiment topic , the probability of which is greater than all the other sentiment topics. Prior information is provided in the form of seed words to FACTS, CFACTS and JST and in the form of overall ratings to FACTS-R and CFACTS-R. It is to be noted that, in the case of FACTS-R and CFACTS-R, the rating forms an important input, however, the polarity is what we are trying to predict in this case. So, in case of FACTS-R and CFACTS-R, this experimentation becomes that of semi-supervised document polarity detection i.e first we allow FACTS-R and CFACTS-R to take advantage of the rating as a source of information and then learn the η parameter (this set can be looked at as the training set) and then using the parameters learnt, we estimate the θ_d^s corresponding to each of the documents in the test set and classify the document accordingly. We used five-fold cross validation technique in case of FACTS-R and CFACTS-R models to report the accuracy.

Discussion The results of the binary classification task and the five-class classification task are shown in the table 10. As can be seen from 10, the CFACTS and the CFACTS-R models outperform the other models in both the tasks. The interesting thing to note is that in the five-class task, CFACTS-R model performs slightly better than CFACTS model because of the sentiments being correlated with the rating. Further analysis revealed that in case of the models which use seeding as the prior information, there is no appropriate way to provide the seeds which are *neutral* and hence those models which do not correlate the rating with the sentiments perform poorly compared to their counterparts. The baseline JST also faces problems with distinguishing between the neutral and the non-neutral opinion topics.

Table 10: Accuracy of polarity detection of reviews

Model	Accuracy %	
	2 Class	5 Class
CFACTS-R	83.98	**77.87**
CFACTS	**84.52**	75.02
FACTS-R	78.02	71.76
FACTS	78.19	70.28
JST	78.38	69.39

5.4.2 Predicting the ratings of reviews FACTS-R and CFACTS-R models tie the sentiments expressed in the review documents with the overall ratings. This enables these models to predict the ratings associated with unrated reviews. sLDA [13] is another such topic model which models the review rating as a response variable. The difference between sLDA and FACTS-R model is that FACTS-R model also incorporates the notion of syntax in order to distinguish between the words which actually correspond to the opinion words and those which do not and model the response variable only as a function of the opinion words. CFACTS-R model goes one more step ahead and brings in the notions of facet and sentiment coherence along with modelling syntactic differences. We tried to compare our models with sLDA in order to see how effective capturing these syntactic differences and coherence prove to be. We evaluate the performance of our models using the predictive R^2 statistic :

CFACTS-R - 0.526, FACTS-R - 0.452, sLDA - 0.468.

On analyzing the results, we found that CFACTS-R was tying together the notions of facets and sentiments and hence was able to model the rating well, however since FACTS-R models the rating only as a function of explicit sentiment words, it is performing poorly compared

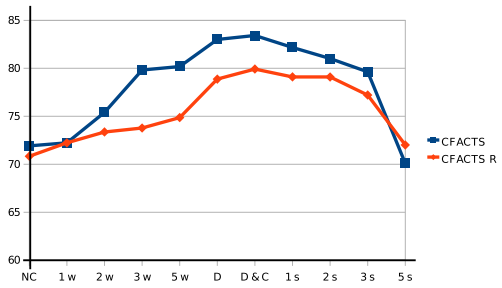


Figure 2: Accuracies of (facet,sentiment) pairs extraction (sentence level)

to CFACTS-R and sLDA.

Impact of Window Size Recall that the coherent models, CFACTS and CFACTS-R, involve windows as the basic unit of coherence. Here, we briefly study the effect of window sizes on facet-sentiment pair extraction. Figure 2 plots this accuracy for all the models at a sentence level for varying window sizes (Note that in the plot, NC corresponds to the 'No-coherence' version of the model with window length of 1 word) Recall that FACTS can be interpreted to have no coherence with a window length of 1 word. We varied window length from a few words (w) , to a few sentences (s) , also considering splits using delimiters (D) (comma, semicolon,stop) and conjunctions (C).

It can be seen that mid-size windows extracted using delimiters, or those that are 1-2 sentences long achieve the best accuracy, while very small and very large window sizes are detrimental. While learning the right window size is a possibility, the results show that coherence can be satisfactorily captured with fixed window sizes of reasonable length.

6 Conclusions

In this paper, we have proposed probabilistic models for facet-based sentiment analysis, which tackle all the aspects of the problem, including discovering latent facets, the sentiment categories and their polarity levels, and the association between facets and sentiments, in a language and domain independent manner without expert intervention. The models jointly capture syntax and semantics, and importantly different notions of coherence in reviews, for performing this task. Using extensive experiments over real world review data, we have demonstrated that the models outperform various state-of-the-art baselines in all aspects of this task.

7 Acknowledgements

CB would like to thank SAP India Ltd. for partially supporting this research. All the authors would like to thank the anonymous reviewers for their valuable suggestions.

References

[1] M. Hu and B. Liu, *Mining and Summarizing Customer Reviews*, Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[2] A. Popescu and O. Etzioni, *Extracting product features and opinions from reviews*, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005.

[3] W. Jin and H. H. Ho, *A novel lexicalized HMM-based learning framework for web opinion mining*, Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

[4] I. Titov and R. McDonald, *A Joint Model of Text and Aspect Ratings for Sentiment Summarization*, 46th Meeting of Association for Computational Linguistics (ACL-08), 2008.

[5] I. Titov and R. McDonald, *Modeling Online Reviews with Multi-Grain Topic Models*, Proceedings of the 17th International World Wide Web Conference (WWW-2008), 2008.

[6] C. Lin and Y. He, *Joint sentiment/topic model for sentiment analysis*, Proceeding of the 18th ACM conference on Information and knowledge management, 2009.

[7] Q. Mei, X. Ling, M. Wondra, H. Su and C. Zhai, *Topic sentiment mixture: modeling facets and opinions in weblogs*, Proceedings of the 16th international conference on World Wide Web, 2007.

[8] T. L. Griffiths and M. Steyvers and D. M. Blei and J. B. Tenenbaum, *Integrating Topics and Syntax*, Advances in Neural Information Processing Systems 17, 2005.

[9] D. M. Blei, M. Jordan and A. Ng, *Latent Dirichlet Allocation*, Journal of Machine Learning and Research, page 993-1022, 2003.

[10] Miller, George A. *WordNet - About Us* WordNet. Princeton University. 2009. http://wordnet.princeton.edu

[11] K. Toutanova, D. Klein, C. Manning, W. Morgan, A. Rafferty, and M. Galley, *Stanford Log-linear Part-Of-Speech Tagger*, 2009. http://nlp.stanford.edu/software/tagger.shtml

[12] T. Hofmann, *Probabilistic latent semantic analysis*, Proceedings of Uncertainty in Artificial Intelligence, Stockholm, 1999.

[13] D. M. Blei and J. D. McAullife, *Supervised topic models*, Neural Information Processing Sytems, 2007.

[14] S. Brody and N. Elhadad, *An Unsupervised Aspect-Sentiment Model for Online Reviews*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 2010.