Modeling Review Comments

Arjun Mukherjee
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
arjun4787@gmail.com

Bing Liu
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
liub@cs.uic.edu

Abstract

Writing comments about news articles, blogs, or reviews have become a popular activity in social media. In this paper, we analyze reader comments about reviews. Analyzing review comments is important because reviews only tell the experiences and evaluations of reviewers about the reviewed products or services. Comments, on the other hand, are readers' evaluations of reviews, their questions and concerns. Clearly, the information in comments is valuable for both future readers and brands. This paper proposes two latent variable models to simultaneously model and extract these key pieces of information. The results also enable classification of comments accurately. Experiments using Amazon review comments demonstrate the effectiveness of the proposed models.

1. Introduction

Online reviews enable consumers to evaluate the products and services that they have used. These reviews are also used by other consumers and businesses as a valuable source of opinions.

However, reviews only give the evaluations and experiences of the reviewers. Often a reviewer may not be an expert of the product and may misuse the product or make other mistakes. There may also be aspects of the product that the reviewer did not mention but a reader wants to know. Some reviewers may even write fake reviews to promote some products, which is called *opinion spamming* (Jindal and Liu 2008). To improve the online review system and user experience, some review hosting sites allow readers to write comments about reviews (apart from just providing a feedback by clicking whether the review is helpful or not). Many reviews receive a large number of comments. It is difficult for a reader to read them to get a gist of them. An automated comment analysis would be very helpful. Review comments mainly contain the following information:

Thumbs-up or thumbs-down: Some readers may comment on whether they find the review useful in helping them make a buying decision.

Agreement or disagreement: Some readers who comment on a review may be users of the product themselves. They often state whether they agree or disagree with the review. Such comments are valuable as they provide a second opinion, which may even identify fake reviews because a genuine user often can easily spot reviewers who have never used the product.

Question and answer: A commenter may ask for clarification or about some aspects of the product that are not covered in the review.

In this paper, we use statistical modeling to model review comments. Two new generative models are proposed. The first model is called the Topic and Multi-Expression model (TME). It models topics and different types of expressions, which represent different types of comment posts:

1. *Thumbs-up* (e.g., "review helped me")
2. *Thumbs-down* (e.g., "poor review")
3. *Question* (e.g., "how to")

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 320–329,
Jeju, Republic of Korea, 8-14 July 2012.

4. *Answer acknowledgement* (e.g., "thank you for clarifying"). Note that we have no expressions for answers to questions as there are usually no specific phrases indicating that a post answers a question except starting with the name of the person who asked the question. However, there are typical phrases for acknowledging answers, thus *answer acknowledgement* expressions.
5. *Disagreement* (*contention*) (e.g., "I disagree")
6. *Agreement* (e.g., "I agree").

For ease of presentation, we call these expressions the *comment expressions* (or *C-expressions*). TME provides a basic model for extracting these pieces of information and topics. Its generative process separates topics and C-expression types using a switch variable and treats posts as random mixtures over latent topics and C-expression types. The second model, called ME-TME, improves TME by using Maximum-Entropy priors to guide topic/expression switching. In short, the two models provide a principled and integrated approach to simultaneously discover topics and C-expressions, which is the goal of this work. Note that topics are usually product aspects in this work.

The extracted C-expressions and topics from review comments are very useful in practice. First of all, C-expressions enable us to perform more accurate classification of comments, which can give us a good evaluation of the review quality and credibility. For example, a review with many *Disagreeing* and *Thumbs-down* comments is dubious. Second, the extracted C-expressions and topics help identify the key product aspects that people are troubled with in disagreements and in questions. Our experimental results in Section 5 will demonstrate these capabilities of our models.

With these pieces of information, comments for a review can be summarized. The summary may include, but not limited to, the following: (1) percent of people who give the review thumbs-up or thumbs-down; (2) percent of people who agree or disagree (or contend) with the reviewer; (3) contentious (disagreed) aspects (or topics); (4) aspects about which people often have questions.

To the best of our knowledge, there is no reported work on such a fine-grained modeling of review comments. The related works are mainly in sentiment analysis (Pang and Lee, 2008; Liu 2012), e.g., topic and sentiment modeling, review quality prediction and review spam detection. However, our work is different from them. We will compare with them in detail in Section 2.

The proposed models have been evaluated both qualitatively and quantitatively using a large number of review comments from Amazon.com. Experimental results show that both TME and ME-TME are effective in performing their tasks. ME-TME also outperforms TME significantly.

2. Related Work

We believe that this work is the first attempt to model review comments for fine-grained analysis. There are, however, several general research areas that are related to our work.

Topic models such as LDA (Latent Dirichlet Allocation) (Blei et al., 2003) have been used to mine topics in large text collections. There have been various extensions to multi-grain (Titov and McDonald, 2008a), labeled (Ramage et al., 2009), partially-labeled (Ramage et al., 2011), constrained (Andrzejewski et al., 2009) models, etc. These models produce only topics but not multiple types of expressions together with topics. Note that in labeled models, each document is labeled with one or multiple labels. For our work, there is no label for each comment. Our labeling is on topical terms and C-expressions with the purpose of obtaining some priors to separate topics and C-expressions.

In sentiment analysis, researchers have jointly modeled topics and sentiment words (Lin and He, 2009; Mei et al., 2007; Lu and Zhai, 2008; Titov and McDonald, 2008b; Lu et al., 2009; Brody and Elhadad, 2010; Wang et al., 2010; Jo and Oh, 2011; Maghaddam and Ester, 2011; Sauper et al., 2011; Mukherjee and Liu, 2012a). Our model is more related to the ME-LDA model in (Zhao et al., 2010), which used a switch variable trained with Maximum-Entropy to separate topic and sentiment words. We also use such a variable. However, unlike sentiments and topics in reviews, which are emitted in the same sentence, C-expressions often interleave with topics across sentences and the same comment post may also have multiple types of C-expressions. Additionally, C-expressions are mostly phrases rather than individual words. Thus, a different model is required to model them.

There have also been works aimed at putting authors in debate into support/oppose camps, e.g., (Galley et al., 2004; Agarwal et al., 2003; Murakami and Raymond, 2010), modeling debate discussions considering reply relations (Mukherjee and Liu, 2012b), and identifying stances in debates (Somasundaran and Wiebe, 2009; Thomas et al.,

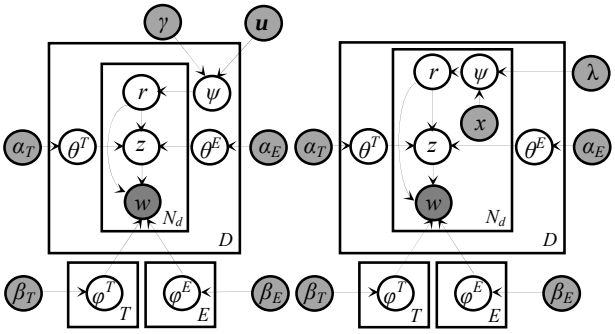
2006; Burfoot et al., 2011). (Yano and Smith, 2010) also modeled the relationship of a blog post and the number of comments it receives. These works are different as they do not mine C-expressions or discover the points of contention and questions in comments.

In (Kim et al., 2006; Zhang and Varadarajan, 2006; Ghose and Ipeirotis, 2007; Liu et al., 2007; Liu et al., 2008; O'Mahony and Smyth, 2009; Tsur and Rappoport 2009), various classification and regression approaches were taken to assess the quality of reviews. (Jindal and Liu, 2008; Lim et al., 2010; Li et al. 2011; Ott et al., 2011; Mukherjee et al., 2012) detect fake reviews and reviewers. However, all these works are not concerned with review comments.

3. The Basic TME Model

This section discusses TME. The next section discusses ME-TME, which improves TME. These models belong to the family of generative models for text where words and phrases (*n*-grams) are viewed as random variables, and a document is viewed as a bag of *n*-grams and each *n*-gram takes a value from a predefined vocabulary. In this work, we use up to 4-grams, i.e., n = 1, 2, 3, 4. For simplicity, we use *terms* to denote both *words* (unigrams or 1-grams) and *phrases* (*n*-grams). We denote the entries in our vocabulary by $v_{1\ldots V}$ where V is the number of unique terms in the vocabulary. The entire corpus contains $d_{1\ldots D}$ documents. A document (e.g., comment post) d is represented as a vector of terms $\boldsymbol{w_d}$ with N_d entries. W is the set of all observed terms with cardinality, $|W| = \sum_d N_d$.

The TME (Topic and Multi-Expression) model is a hierarchical generative model motivated by the joint occurrence of various types of expressions indicating *Thumbs-up*, *Thumbs-down*, *Question*, *Answer acknowledgement*, *Agreement,* and *Disagreement* and topics in comment posts. As before, these expressions are collectively called C-expressions. A typical comment post mentions a few topics (using semantically related topical terms) and expresses some viewpoints with one or more C-expression types (using semantically related expressions). This observation motivates the generative process of our model where documents (posts) are represented as random mixtures of latent topics and C-expression types. Each topic or C-expression type is characterized by a distribution over terms (words/phrases). Assume



(a) TME Model (b) ME-TME Model

Figure 1: Graphical Models in plate notations.

we have $t_{1\ldots T}$ topics and $e_{1\ldots E}$ expression types in our corpus. Note that in our case of Amazon review comments, based on reading various posts, we hypothesize that E = 6 as in such review discussions, we mostly find 6 expression types (more details in Section 5.1). Let ψ_d denote the distribution of topics and C-expressions in a document d with $r_{d,j} \in \{\hat{t}, \hat{e}\}$ denoting the binary indicator variable (topic or C-expression) for the j^{th} term of d, $w_{d,j}$. $z_{d,j}$denotes the appropriate topic or C-expression type index to which $w_{d,j}$ belongs. We parameterize multinomials over topics using a matrix $\Theta^T_{D\times T}$whose elements $\theta^T_{d,t}$ signify the probability of document d exhibiting topic t. For simplicity of notation, we will drop the latter subscript (t in this case) when convenient and use θ^T_d to stand for the d^{th} row of Θ^T. Similarly, we define multinomials over C-expression types using a matrix $\Theta^E_{D\times E}$. The multinomials over terms associated with each topic are parameterized by a matrix $\Phi^T_{T\times V}$, whose elements $\varphi^T_{t,v}$ denote the probability of generating v from topic t. Likewise, multinomials over terms associated with each C-expression type are parameterized by a matrix $\Phi^E_{E\times V}$. We now define the generative process of TME (see Figure 1(a)).

A. For each C-expression type e, draw $\varphi^E_e \sim Dir(\beta_E)$
B. For each topic t, draw $\varphi^T_t \sim Dir(\beta_T)$
C. For each comment post $d \in \{1 \ldots D\}$:
   i. Draw $\psi_d \sim Beta(\gamma \boldsymbol{u})$
   ii. Draw $\theta^E_d \sim Dir(\alpha_E)$
   iii. Draw $\theta^T_d \sim Dir(\alpha_T)$
   iv. For each term $w_{d,j}$, $j \in \{1 \ldots N_d\}$:
      a. Draw $r_{d,j} \sim Bernoulli(\psi_d)$
      b. if ($r_{d,j} = \hat{e}$ // $w_{d,j}$is a C-expression term
            Draw $z_{d,j} \sim Mult(\theta^E_d)$
         else // $r_{d,j} = \hat{t}$, $w_{d,j}$is a topical term
            Draw $z_{d,j} \sim Mult(\theta^T_d)$
      c. Emit $w_{d,j} \sim Mult(\varphi^{r_{d,j}}_{z_{d,j}})$

To learn the TME model from data, as exact inference is not possible, we resort to approximate inference using *collapsed Gibbs sampling* (Griffiths and Steyvers, 2004). Gibbs sampling is a form of Markov Chain Monte Carlo method where a Markov chain is constructed to have a particular stationary distribution. In our case, we want to construct a Markov chain which converges to the posterior distribution over R and Z conditioned on the data. We only need to sample z and r as we use collapsed Gibbs sampling and the dependencies of θ and φ have been integrated out analytically in the joint. Denoting the random variables $\{w, z, r\}$ by singular subscripts$\{w_k, z_k, r_k\}$, $k_{1\ldots K}$, where $K = \sum_d N_d$, a single iteration consists of performing the following sampling:

$$p(z_k = t, r_k = \hat{t} | W_{\neg k}, Z_{\neg k}, R_{\neg k}, w_k = v) \propto$$
$$\frac{n^T_{d\,\neg k} + \gamma_a}{n^T_{d\,\neg k} + n^E_{d\,\neg k} + \gamma_a + \gamma_b} \times \frac{n^{DT}_{d,t\,\neg k} + \alpha_T}{n^{DT}_{d,(\cdot)\,\neg k} + T\alpha_T} \times \frac{n^{CT}_{t,v\,\neg k} + \beta_T}{n^{CT}_{t,(\cdot)\,\neg k} + V\beta_T} \quad (1)$$

$$p(z_k = e, r_k = \hat{e} | W_{\neg k}, Z_{\neg k}, R_{\neg k}, w_k = v) \propto$$
$$\frac{n^E_{d\,\neg k} + \gamma_b}{n^T_{d\,\neg k} + n^E_{d\,\neg k} + \gamma_a + \gamma_b} \times \frac{n^{DE}_{d,e\,\neg k} + \alpha_E}{n^{DE}_{d,(\cdot)\,\neg k} + E\alpha_E} \times \frac{n^{CE}_{e,v\,\neg k} + \beta_E}{n^{CE}_{e,(\cdot)\,\neg k} + V\beta_E} \quad (2)$$

where $k = (d, j)$ denotes the j^{th} term of document d and the subscript $\neg k$ denotes assignments excluding the term at (d, j). Counts$n^{CT}_{t,v}$ and $n^{CE}_{e,v}$ denote the number of times term v was assigned to topic t and expression type e respectively. $n^{DT}_{d,t}$ and $n^{DE}_{d,e}$ denote the number of terms in document d that were assigned to topic t and C-expression type e respectively. Lastly, n^T_d and n^E_d are the number of terms in d that were assigned to topics and C-expression types respectively. Omission of the latter index denoted by $(\cdot)$ represents the marginalized sum over the latter index. We employ a blocked sampler jointly sampling r and z as this improves convergence and reduces autocorrelation of the Gibbs sampler (Rosen-Zvi et al., 2004).

Asymmetric Beta priors: Based on our initial experiments with TME, we found that properly setting the smoothing hyper-parameter $\gamma\boldsymbol{u}$ is crucial as it governs the topic/expression switch.

According to the generative process, ψ_d is the (success) probability (of the Bernoulli distribution) of emitting a topical/aspect term in a comment post d and$1 - \psi_d$, the probability of emitting a C-expression term in d. Without loss of generality, we draw $\psi_d \sim Beta(\gamma\boldsymbol{u})$ where γ is the concentration parameter and $\boldsymbol{u} = [u_a, u_b]$ is the base measure. Without any prior belief, one resorts to uniform base measure $u_a = u_b = 0.5$ (i.e., assumes that both topical and C-expression terms are equally likely to be emitted in a comment post). This results in symmetric Beta priors $\psi_d \sim Beta(\gamma_a, \gamma_b)$ where $\gamma_a = \gamma u_a$, $\gamma_b = \gamma u_b$ and $\gamma_a = \gamma_b = \gamma/2$. However, knowing the fact that topics are more likely to be emitted than expressions in a post *apriori* motivates us to take guidance from asymmetric priors (i.e., we now have a non-uniform base measure$\boldsymbol{u}$). This asymmetric setting of γ ensures that samples of ψ_d are more close to the actual distribution of topical terms in posts based on some domain knowledge. Symmetric γ cannot utilize any prior knowledge. In (Lin and He, 2009), a method was proposed to incorporate domain knowledge during Gibbs sampling initialization, but its effect becomes weak as the sampling progresses (Jo and Oh, 2011).

For asymmetric priors, we estimate the hyper-parameters from labeled data. Given a labeled set D_L, where we know the per post probability of C-expression emission $(1 - \psi_d)$, we use the method of moments to estimate $\gamma = [\gamma_a, \gamma_b]$ as follows:

$$\gamma_a = \mu\left(\frac{\mu(1-\mu)}{\sigma} - 1\right), \gamma_b = \gamma_a\left(\frac{1}{\mu} - 1\right);\ \mu = E[\psi_d], \sigma = Var[\psi_d] \quad (3)$$

4. ME-TME Model

The guidance of Beta priors, although helps, is still relatively coarse and weak. We can do better to produce clearer separation of topical and C-expression terms. An alternative strategy is to employ *Maximum-Entropy* (Max-Ent) priors instead of Beta priors. The Max-Ent parameters can be learned from a small number of labeled topical and C-expression terms (words and phrases) which can serve as good priors. The idea is motivated by the following observation: topical and C-expression terms typically play different syntactic roles in a sentence. Topical terms (e.g. "ipod" "cell phone", "macro lens", "kindle", etc.) tend to be noun and noun phrases while expression terms ("I refute", "how can you say", "great review") usually contain pronouns, verbs, wh-determiners, adjectives, and modals. In order to utilize the part-of-speech (POS) tag information, we move the topic/C-expression distribution ψ_d (the prior over the indicator variable $r_{d,j}$) from the document plate to the word plate (see Figure 1 (b)) and draw it from a Max-Ent model conditioned on the observed feature vector $\overrightarrow{x_{d,j}}$ associated with $w_{d,j}$ and the learned Max-Ent parameters λ. $x_{d,j}$ can

encode arbitrary contextual features for learning. With Max-Ent priors, we have the new model ME-TME. In this work, we encode both lexical and POS features of the previous, current and next POS tags/lexemes of the term $w_{d,j}$. More specifically,

$$\overrightarrow{x_{d,j}} = \left[POS_{w_{d,j}-1}, POS_{w_{d,j}}, POS_{w_{d,j}+1}, w_{d,j}-1, w_{d,j}, w_{d,j}+1\right]$$

For phrasal terms (n-grams), all POS tags and lexemes of $w_{d,j}$are considered as features. Incorporating Max-Ent priors, the Gibbs sampler of ME-TME is given by:

$$p(z_k = t, r_k = \hat{t} | W_{\neg k}, Z_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, \hat{t}))}{\sum_{y \in \{\hat{e}, \hat{t}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, y))} \times \frac{n_{d,t}^{DT}{}_{\neg k} + \alpha_T}{n_{d,(\cdot)}^{DT}{}_{\neg k} + T\alpha_T} \times \frac{n_{t,v}^{CT}{}_{\neg k} + \beta_T}{n_{t,(\cdot)}^{CT}{}_{\neg k} + V\beta_T} \quad (4)$$

$$p(z_k = e, r_k = \hat{e} | W_{\neg k}, Z_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, \hat{e}))}{\sum_{y \in \{\hat{e}, \hat{t}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, y))} \times \frac{n_{d,e}^{DE}{}_{\neg k} + \alpha_E}{n_{d,(\cdot)}^{DE}{}_{\neg k} + E\alpha_E} \times \frac{n_{e,v}^{CE}{}_{\neg k} + \beta_E}{n_{e,(\cdot)}^{CE}{}_{\neg k} + V\beta_E} \quad (5)$$

where $\lambda_{1\ldots n}$ are the parameters of the learned Max-Ent model corresponding to the n binary feature functions $f_{1\ldots n}$ from Max-Ent.

5. Evaluation

We now evaluate the proposed TME and ME-TME models. Specifically, we evaluate the discovered C-expressions, contentious aspects, and aspects often mentioned in questions.

5.1 Dataset and Experiment Settings We crawled comments of reviews in Amazon.com for a variety of products. For each comment we extracted its id, the comment author id, the review id on which it commented, and the review author id. Our database consisted of 21,316 authors, 37,548 reviews, and 88,345 comments with an average of 124 words per comment post.

For all our experiments, the hyper-parameters for TME and ME-TME were set to the heuristic values α_T = 50/T, α_E = 50/E, β_T = β_E = 0.1 as suggested in (Griffiths and Steyvers, 2004). For γ, we estimated the asymmetric Beta priors using the method of moments discussed in Section 3. We sampled 1000 random posts and for each post we identified the C-expressions emitted. We thus computed the per-post probability of C-expression emission $(1 - \psi_d)$ and used Eq. (3) to get the final estimates, γ_a = 3.66, γ_b= 1.21. To learn the Max-Ent parameters λ, we randomly sampled 500 terms from our corpus appearing at least 10 times and labeled them as topical (332) or C-expressions (168) and used the corresponding feature vector of each term (in the context of posts where it occurs) to train the Max-Ent model. We set the number of topics, T = 100 and the number of C-expression types, E = 6 (*Thumbs-up*, *Thumbs-down*, *Question*, *Answer acknowledgement*, *Agreement* and *Disagreement*) as in review comments, we usually find these six dominant expression types. Note that knowing the exact number of topics, T and expression types, E in a corpus is difficult. While non-parametric Bayesian approaches (Teh et al., 2006) aim to estimate T from the corpus, in this work the heuristic values obtained from our initial experiments produced good results. We also tried increasing E to 7, 8, etc. However, it did not produce any new dominant expression type. Instead, the expression types became less specific as the expression term space became sparser.

5.2 C-Expression Evaluation

We now evaluate the discovered C-expressions. We first evaluate them qualitatively in Tables 1 and 2. Table 1 shows the top terms of all expression types using the TME model. We find that TME can discover and cluster many correct C-expressions, e.g., "great review", "review helped me" in *Thumbs-up*; "poor review", "very unfair review" in *Thumbs-down*; "how do I", "help me decide" in *Question*; "good reply", "thank you for clarifying" in *Answer Acknowledgement*; "I disagree", "I refute" in *Disagreement*; and "I agree", "true in fact" in *Agreement*. However, with the guidance of Max-Ent priors, ME-TME did much better (Table 2). For example, we find "level headed review", "review convinced me" in *Thumbs-up*; "biased review", "is flawed" in *Thumbs-down*; "any clues", "I was wondering how" in *Question*; "clears my", "valid answer" in *Answer-acknowledgement*; "I don't buy your", "sheer nonsense" in *Disagreement*; "agree completely", "well said" in *Agreement*. These newly discovered phrases by ME-TME are marked in *blue* in Table 3. ME-TME also has fewer errors.

Next, we evaluate them quantitatively using the metric *precision @ n*, which gives the precision at different rank positions. This metric is appropriate here because the C-expressions (according to top terms in Φ^E) produced by TME and ME-TME are rankings. Table 3 reports the precisions @ top 25, 50, 75, and 100 rank positions for all six expression types across both models. We evaluated till top 100 positions because it is usually

Thumbs-up (e_1): **review**, **thanks**, great review, nice review, **time**, best review, **appreciate**, **you**, your review helped, nice, terrific, review helped me, good critique, **very**, **assert**, **wrong**, useful review, **don't**, **misleading**, thanks a lot, …

Thumbs-down (e_2): **review**, **no**, poor review, imprecise, **you**, **complaint**, **very**, suspicious, bogus review, **absolutely**, **credible**, very unfair review, criticisms, **true**, disregard this review, disagree with, **judgment**, without owning, …

Question (e_3): question, **my**, **I**, how do I, why isn't, please explain, **good answer**, clarify, don't understand, my doubts, I'm confused, **does not**, **understand**, help me decide, how to, **yes**, **answer**, how can I, **can't explain**, …

Answer Acknowledgement (e_4): **my**, **informative**, answer, good reply, thank you for clarifying, answer doesn't, good answer, **vague**, helped me choose, useful suggestion, **don't understand**, **cannot explain**, your answer, **doubts**, answer isn't, …

Disagreement (e_5): disagree, **I**, **don't**, I disagree, argument claim, I reject, I refute, I refuse, oppose, debate, **accept**, don't agree, **quote**, **sense**, would disagree, **assertions**, I doubt, **right**, your, **really**, you, I'd disagree, cannot, nonsense,...

Agreement (e_6): yes, **do**, correct, indeed, **no**, right, I agree, **you**, agree, I accept, **very**, yes indeed, true in fact, indeed correct, I'd agree, **completely**, true, **but**, **doesn't**, **don't**, definitely, **false**, completely agree, agree with your, true, …

Table 1: Top terms (comma delimited) of six expression types e_1, e_2, e_3, e_4, e_5, e_6 (Φ^E) using TME model. **Red** (**bold**) colored terms denote possible errors

Thumbs-up (e_1): **review**, **you**, great review, *I'm glad I read,* best review, *review convinced me*, review helped me, good review, *terrific review*, **job**, *thoughtful review*, awesome review, *level headed review*, *good critique*, good job, *video review*,...

Thumbs-down (e_2): **review**, **you**, bogus review, con, useless review, *ridiculous*, *biased review*, very unfair review, *is flawed*, **completely**, *skeptical*, *badmouth*, *misleading review*, cynical review, wrong, disregard this review, *seemingly honest*, …

Question (e_3): question, **I**, how do I, why isn't, please explain, clarify, *any clues*, **answer**, please explain, help me decide, **vague**, how to, how do I, where can I, *how to set*, *I was wondering how*, *could you explain*, how can I, *can I use*, …

Answer Acknowledgement (e_4): **my**, good reply, , answer, reply, helped me choose, *clears my,* *valid answer*, answer doesn't, *satisfactory answer*, **can you clarify**, *informative answer*, useful suggestion, *perfect answer*, thanks for your reply, **doubts**, …

Disagreement (e_5): disagree, **I**, **don't**, I disagree, doesn't, *I don't buy your*, **credible**, I reject, I doubt, I refuse, I *oppose*, *sheer nonsense*, **hardly**, don't agree, *can you prove*, *you have no clue*, *how do you say*, **sense**, *you fail*, *contradiction*, …

Agreement (e_6): **I**, **do**, agree, **point**, yes, **really**, *would agree*, **you**, agree, I accept, **claim**, *agree completely*, *personally agree*, true in fact, indeed correct, *well said*, *valid point*, correct, **never meant**, **might not**, *definitely agree*,…

Table 2: Top terms (comma delimited) of six expression types using ME-TME model. **Red** (**bold**) terms denote possible errors. *Blue* (*italics*) terms denote those newly discovered by the model; rest (black) were used in Max-Ent training.

important to see whether a model can discover and rank those major expressions of a type at the top. We believe that top 100 are sufficient for most applications. From Table 3, we observe that ME-TME consistently outperforms TME in precisions across all expression types and all rank positions. This shows that Max-Ent priors are more effective in discovering expressions than Beta priors. Note that we couldn't compare with an existing baseline because there is no reported study on this problem.

5.3 Comment Classification

Here we show that the discovered C-expressions can help comment classification. Note that since a comment can belong to one or more types (e.g., a comment can belong to both *Thumbs-up* and A*greement* types), this task is an instance of multi-label classification, i.e., an instance can have more than one class label. In order to evaluate all the expression types, we follow the binary approach which is an extension of one-against-all method for multi-label classification. Thus, for each label, we build a binary classification problem. Instances associated with that label are in one class and the rest are in the other class. To perform this task, we randomly sampled 2000 comments, and labeled each of them into one or more of the following 8 labels: *Thumbs-up*, *Thumbs-down*, *Disagreement*, *Agreement*, *Question*, *Answer-Acknowledgement*, *Answer*, and *None*, which have 432, 401, 309, 276, 305, 201, 228, and 18 comments respectively. We disregard the *None* category due to its small size. This labeling is a fairly easy task as one can almost certainly make out to which type a comment belongs. Thus we didn't use multiple labelers. The distribution reveals that the labels are overlapping. For instance, we found many comments belonging to both *Thumbs-down* and *Disagreement*, *Thumbs-up* with *Acknowledgement* and with *Question*.

For supervised classification, the choice of feature is a key issue. While word and POS n-grams are traditional features, such features may not be the best for our task. We now compare such features with the C-expressions discovered by the proposed models. We used the top 1000 terms from each of the 6 C-expression rankings as features. As comments in *Question* type mostly use the punctuation "?", we added it in our feature set. We use precision, recall and F_1 as our metric to compare classification performance using a trained SVM (linear kernel). All results (Table 4) were computed using 10-fold cross-validation (CV). We also tried Naïve Bayes and Logistic Regression classifiers, but they were poorer than SVM. Hence their results are not reported due to space constraints. As a separate experiment (not shown here also due to space constraints), we analyzed the classification performance by varying the number of top terms from 200, 400,…, 1000, 1200, etc. and found that the F_1 scores stabilized after top

C-Expression Type	P@25		P@50		P@75		P@100	
	TME	ME-TME	TME	ME-TME	TME	ME-TME	TME	ME-TME
Thumbs-up	0.60	0.80	0.66	0.78	0.60	0.69	0.55	0.64
Thumbs-down	0.68	0.84	0.70	0.80	0.63	0.67	0.60	0.65
Question	0.64	0.80	0.68	0.76	0.65	0.72	0.61	0.67
Answer-Acknowledgement	0.68	0.76	0.62	0.72	0.57	0.64	0.54	0.58
Disagreement	0.76	0.88	0.74	0.80	0.68	0.73	0.65	0.70
Agreement	0.72	0.80	0.64	0.74	0.61	0.70	0.60	0.69

Table 3: Precision @ top 25, 50, 75, and 100 rank positions for all C-expression types.

Features	Thumbs-up			Thumbs-down			Question			Answer-Ack.			Disagreement			Agreement			Answer		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
W+POS 1-gram	0.68	0.66	0.67	0.65	0.65	0.65	0.71	0.68	0.69	0.64	0.61	0.62	0.73	0.72	0.72	0.67	0.65	0.66	0.58	0.57	0.57
W+POS 1-2 gram	0.72	0.69	0.70	0.68	0.67	0.67	0.74	0.69	0.71	0.69	0.63	0.65	0.76	0.75	0.75	0.71	0.69	0.70	0.60	0.57	0.58
W+POS, 1-3 gram	0.73	0.71	0.72	0.69	0.68	0.68	0.75	0.69	0.72	0.70	0.64	0.66	0.76	0.76	0.76	0.72	0.70	0.71	0.61	0.58	0.59
W+POS, 1-4 gram	0.74	0.72	0.73	0.71	0.68	0.69	0.75	0.70	0.72	0.70	0.65	0.67	0.77	0.76	0.76	0.73	0.70	0.71	0.61	0.58	0.59
C-Expr. Φ^E, TME	0.82	0.74	0.78	0.77	0.71	0.74	0.83	0.75	0.78	0.75	0.72	0.73	0.83	0.80	0.81	0.78	0.75	0.76	0.66	0.61	0.63
C-Expr. Φ^E, ME-TME	0.87	0.79	0.83	0.80	0.73	0.76	0.87	0.76	0.81	0.77	0.72	0.74	0.86	0.81	0.83	0.81	0.77	0.79	0.67	0.61	0.64

Table 4: Precision (P), Recall (R), and F_1 scores of binary classification using SVM and different features. The improvements of our models are significant (p<0.001) over paired t-test across 10-fold cross validation.

D	Φ^E + Noun/Noun Phrase						TME						ME-TME					
	J_1			J_2			J_1			J_2			J_1			J_2		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
D1	0.62	0.70	0.66	0.58	0.67	0.62	0.66	0.75	0.70	0.62	0.70	0.66	0.67	0.79	0.73	0.64	0.74	0.69
D2	0.61	0.67	0.64	0.57	0.63	0.60	0.66	0.72	0.69	0.62	0.67	0.64	0.68	0.75	0.71	0.64	0.71	0.67
D3	0.60	0.69	0.64	0.56	0.64	0.60	0.64	0.73	0.68	0.60	0.67	0.63	0.67	0.76	0.71	0.63	0.72	0.67
D4	0.59	0.68	0.63	0.55	0.65	0.60	0.63	0.71	0.67	0.59	0.68	0.63	0.65	0.73	0.69	0.62	0.71	0.66
Avg.	0.61	0.69	0.64	0.57	0.65	0.61	0.65	0.73	0.69	0.61	0.68	0.64	0.67	0.76	0.71	0.63	0.72	0.67

Table 5 (a)

D	Φ^E + Noun/Noun Phrase						TME						ME-TME					
	J_1			J_2			J_1			J_2			J_1			J_2		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
D1	0.57	0.65	0.61	0.54	0.63	0.58	0.61	0.69	0.65	0.58	0.66	0.62	0.64	0.73	0.68	0.61	0.70	0.65
D2	0.61	0.66	0.63	0.58	0.61	0.59	0.64	0.68	0.66	0.60	0.64	0.62	0.68	0.70	0.69	0.65	0.69	0.67
D3	0.60	0.68	0.64	0.57	0.64	0.60	0.64	0.71	0.67	0.62	0.68	0.65	0.67	0.72	0.69	0.64	0.69	0.66
D4	0.56	0.67	0.61	0.55	0.65	0.60	0.60	0.72	0.65	0.58	0.68	0.63	0.63	0.75	0.68	0.61	0.71	0.66
Avg.	0.59	0.67	0.62	0.56	0.63	0.59	0.62	0.70	0.66	0.60	0.67	0.63	0.66	0.73	0.69	0.63	0.70	0.66

Table 5 (b)

Table 5: Points of Contention (a), Questioned aspects (b). D1: Ipod, D2: Kindle, D3: Nikon, D4: Garmin. We report the average precision (P), recall (R), and F_1 score over 100 comments for each particular domain.

Statistical significance: Differences between Nearest Noun Phrase and TME for both judges (J_1, J_2) across all domains were significant at 97% confidence level (p<0.03). Differences among TME and ME-TME for both judges (J_1, J_2) across all domains were significant at 95% confidence level (p<0.05). A paired t-test was used for testing significance.

1000 terms. From Table 4, we see that F_1 scores dramatically increase with C-expression (Φ^E) features for all expression types. TME and ME-TME progressively improve the classification. Improvements of TME and ME-TME being significant (p<0.001) using a paired t-test across 10-fold cross validations shows that the discovered C-expressions are of high quality and useful.

We note that the annotation resulted in a new label "Answer" which consists of mostly replies to comments with questions. Since an "answer" to a question usually does not show any specific expression, it does not attain very good F_1 scores. Thus, to improve the performance of the *Answer* type comments, we added three binary features for each comment c on top of C-expression features:

i) Is the author of c the review author too? The idea here is that most of the times the reviewer answers the questions raised in comments.
ii) Is there any comment posted before c by some author a which has been previously classified as a question post?
iii) Is there any comment posted after c by author a that replies to c (using @name) and is an *Answer-Acknowledgement* comment (which again has been previously classified as such)?

Using these additional features, we obtained a precision of 0.78 and a recall of 0.73 yielding an F_1

score of 0.75 which is a dramatic increase beyond 0.64 achieved by ME-TME in Table 4.

5.4 Contention Points and Questioned Aspects

We now turn to the task of discovering points of contention in disagreement comments and aspects (or topics) raised in questions. By "points", we mean the topical terms on which some contentions or disagreements have been expressed. Topics being the product aspects are also indirectly evaluated in this task. We employ the TME and ME-TME models in the following manner.

We only detail the approach for disagreement comments. The same method is applied to question comments. Given a disagreement comment post d, we first select the top k topics that are mentioned in d according to its topic distribution, θ_d^T. Let T_d be the set of these top k topics in d. Then, for each disagreement expression $w \in d \cap \varphi^E_{e=Disagreement}$, we emit the topical terms (words/phrases) of topics in T_d which appear within a word window of q from w in d. More precisely, we emit the set $A = \{w | w \in d \cap \varphi_t^T, t \in T_d, |posi(w) - posi(v)| \leq q\}$, where $posi(\cdot)$ returns the position index of the word or phrase in document d. To compute the intersection $w \in d \cap \varphi_t^T$, we need a threshold. This is so because the Dirichlet distribution has a smoothing effect which assigns some non-zero probability mass to every term in the vocabulary for each topic t. So for computing the intersection, we considered only terms in φ_t^T which have $p(v|t) = \varphi_{t,v}^T > 0.001$ as probability masses lower than 0.001 are more due to the smoothing effect of the Dirichlet distribution than true correlation. In an actual application, the values for k and q can be set according to the user's need. In our experiment, we used $k = 3$ and $q = 5$, which are reasonable because a post normally does not talk about many topics (k), and the contention points (aspect terms) appear quite close to the disagreement expressions.

For comparison, we also designed a baseline. For each disagreement (or question) expression $w \in d \cap \varphi^E_{e=Disagreement}(\varphi^E_{e=Question})$, we emit the nouns and noun phrases within the same window q as the points of contention (question) in d. This baseline is reasonable because topical terms are usually nouns and noun phrases and are near disagreement (question) expressions. We note that this baseline cannot stand alone because it has to rely on our expression models Φ^E of ME-TME.

Next, to evaluate the performance of these methods in discovering points of contention, we randomly selected 100 disagreement (contentious) (and 100 question) comment posts on reviews from each of the 4 product domains: Ipod, Kindle, Nikon Cameras, and Garmin GPS in our database and employed the aforementioned methods to discover the points of contention (question) in each post. Then we asked two human judges (graduate students fluent in English) to manually judge the results produced by each method for each post. We asked them to report the precision of the discovered terms for a post by judging them as being indeed valid points of contention and report recall in a post by judging how many of actually contentious points in the post were discovered. In Table 5 (a), we report the average precision and recall for 100 posts in each domain by the two judges J_1 and J_2 for different methods on the task of discovering points (aspects) of contention. In Table 5 (b), similar results are reported for the task of discovering questioned aspects in 100 question comments for each product domain. Since this judging task is subjective, the differences in the results from the two judges are not surprising. Our judges were made to work in isolation to prevent any bias. We observe that across all domains, ME-TME again performs the best consistently. Note that agreement study using Kappa is not used here as our problem is not to label a fixed set of items categorically by the judges.

6. Conclusion

This paper proposed the problem of modeling review comments, and presented two models TME and ME-TME to model and to extract topics (aspects) and various comment expressions. These expressions enable us to classify comments more accurately, and to find contentious aspects and questioned aspects. These pieces of information also allow us to produce a simple summary of comments for each review as discussed in Section 1. To our knowledge, this is the first attempt to analyze comments in such details. Our experiments demonstrated the efficacy of the models. ME-TME also outperformed TME significantly.

Acknowledgments

This work is supported in part by National Science Foundation (NSF) under grant no. IIS-1111092.

References

Agarwal, R., S. Rajagopalan, R. Srikant, Y. Xu. 2003. Mining newsgroups using networks arising from social behavior. Proceedings of International Conference on World Wide Web 2003.

Andrzejewski, D., X. Zhu, M. Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. Proceedings of International Conference on Machine Learning.

Blei, D., A. Ng, and M. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research.

Brody, S. and S. Elhadad. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. Proceedings of the Annual Conference of the North American Chapter of the ACL.

Burfoot, C., S. Bird, and T. Baldwin. 2011. Collective Classification of Congressional Floor-Debate Transcripts. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

Galley, M., K. McKeown, J. Hirschberg, E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. Proceedings of the 42th Annual Meeting of the Association of Computational Linguistics.

Ghose, A. and P. Ipeirotis. 2007. Designing novel review ranking systems: predicting the usefulness and impact of reviews. Proceedings of International Conference on Electronic Commerce.

Griffiths, T. and M. Steyvers. 2004. Finding scientific topics. Proceedings of National Academy of Sciences.

Kim, S., P. Pantel, T. Chklovski, and M. Pennacchiotti. 2006. Automatically assessing review helpfulness. Proceedings of Empirical Methods in Natural Language Processing.

Jindal, N. and B. Liu. 2008. Opinion spam and analysis. Proceedings of the ACM International Conference on Web Search and Web Data Mining.

Jo, Y. and A. Oh. 2011. Aspect and sentiment unification model for online review analysis. Proceedings of the ACM International Conference on Web Search and Web Data Mining.

Li, F., M. Huang, Y. Yang, and X. Zhu. 2011. Learning to Identify Review Spam. in Proceedings of the International Joint Conference on Artificial Intelligence.

Lim, E., V. Nguyen, N. Jindal, B. Liu, and H. Lauw. 2010. Detecting Product Review Spammers using Rating Behaviors. Proceedings of the ACM International Conference on Information and Knowledge Management.

Lin, C. and Y. He. 2009. Joint sentiment/topic model for sentiment analysis. Proceedings of the ACM International Conference on Information and Knowledge Management.

Liu, J., Y. Cao, C. Lin, Y. Huang, and M. Zhou. 2007. Low-quality product review detection in opinion summarization. Proceedings of Empirical Methods in Natural Language Processing.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool publishers (to appear in June 2012).

Liu, Y., X. Huang, A. An, and X. Yu. 2008. Modeling and predicting the helpfulness of online reviews. Proceedings of IEEE International Conference on Data Mining.

Lu, Y. and C. Zhai. 2008. Opinion integration through semi-supervised topic modeling. Proceedings of International Conference on World Wide Web.

Lu, Y., C. Zhai, and N. Sundaresan. 2009. Rated aspect summarization of short comments. Proceedings of International Conference on World Wide.

Mei, Q. X. Ling, M. Wondra, H. Su and C. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. Proceedings of International Conference on World Wide.

Moghaddam, S. and M. Ester. 2011. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. Proceedings of Annual ACM SIGIR Conference on Research and Development in Information Retrieval.

Mukherjee, A. and B. Liu. 2012a. Aspect Extraction through Semi-Supervised Modeling. Proceedings of 50th Annual Meeting of Association for Computational Linguistics (to appear in July 2012).

Mukherjee, A. and B. Liu. 2012b. Mining Contentions from Discussions and Debates. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (to appear in August 2012).

Mukherjee, A., B. Liu and N. Glance. 2012. Spotting Fake Reviewer Groups in Consumer Reviews. Proceedings of International World Wide Web Conference.

Murakami A., and R. Raymond, 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. Proceedings of International Conference on

Computational Linguistics.

O'Mahony, M. P. and B. Smyth. 2009. Learning to recommend helpful hotel reviews. Proceedings of the third ACM conference on Recommender systems.

Ott, M., Y. Choi, C. Cardie, and J. T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval.

Ramage, D., D. Hall, R. Nallapati, and C. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of Empirical Methods in Natural Language Processing.

Ramage, D., C. Manning, and S. Dumais. 2011 Partially labeled topic models for interpretable text mining. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smith. 2004. The author-topic model for authors and documents. Uncertainty in Artificial Intelligence.

Sauper, C. A. Haghighi and R. Barzilay. 2011. Content models with attitude. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

Somasundaran, S., J. Wiebe. 2009. Recognizing stances in online debates. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP

Teh, Y., M. Jordan, M. Beal and D. Blei. 2006. Hierarchical Dirichlet Processes. Journal of the American Statistical Association.

Thomas, M., B. Pang and L. Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. Proceedings of Empirical Methods in Natural Language Processing.

Titov, I. and R. McDonald. 2008a. Modeling online reviews with multi-grain topic models. Proceedings of International Conference on World Wide Web.

Titov, I. and R. McDonald. 2008b. A joint model of text and aspect ratings for sentiment summarization. Proceedings of Annual Meeting of the Association for Computational Linguistics.

Tsur, O. and A. Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. Proceedings of the International AAAI Conference on Weblogs and Social Media.

Wang, H., Y. Lu, and C. Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Yano, T and N. Smith. 2010. What's Worthy of Comment? Content and Comment Volume in Political Blogs. Proceedings of the International AAAI Conference on Weblogs and Social Media.

Zhang, Z. and B. Varadarajan. 2006. Utility scoring of product reviews. Proceedings of ACM International Conference on Information and Knowledge Management.

Zhao, X., J. Jiang, H. Yan, and X. Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. Proceedings of Empirical Methods in Natural Language Processing.