Complementary Aspect-Based Opinion Mining

Yuan Zuo, Junjie Wu, Hui Zhang, Deqing Wang, and Ke Xu

Abstract—Aspect-based opinion mining is finding elaborate opinions towards a subject such as a product or an event. With explosive growth of opinionated texts on the Web, mining aspect-level opinions has become a promising means for online public opinion analysis. In particular, the boom of various types of online media provides diverse yet complementary information, bringing unprecedented opportunities for cross media aspect-opinion mining. Along this line, we propose CAMEL, a novel topic model for complementary aspect-based opinion mining across asymmetric collections. CAMEL gains information complementarity by modeling both common and specific aspects across collections, while keeping all the corresponding opinions for contrastive study. An auto-labeling scheme called AME is also proposed to help discriminate between aspect and opinion words without elaborative human labeling, which is further enhanced by adding word embedding-based similarity as a new feature. Moreover, CAMEL-DP, a nonparametric alternative to CAMEL is also proposed based on coupled Dirichlet Processes. Extensive experiments on real-world multi-collection reviews data demonstrate the superiority of our methods to competitive baselines. This is particularly true when the information shared by different collections becomes seriously fragmented. Finally, a case study on the public event "2014 Shanghai Stampede" demonstrates the practical value of CAMEL for real-world applications.

Index Terms—Aspect-based opinion mining, latent dirichlet allocation (LDA), maximum entropy model, dirichlet process, word embedding

1 Introduction

With the dramatic growth of opinionated user generated content on the Web, automatically extracting, understanding and summarizing public opinions expressed on online media platforms has become an important research topic and gained much attention in recent years [14], [32]. Aspect-based opinion mining, a technique proposed originally to find detailed opinions towards a perspective of a given product [22], has become a promising challenge for mining aspect-level public opinions from online social media, where the concept of an aspect has been extended to an underlying theme, perspective or viewpoint towards a public event [25], [30], [31]. For instance, for a specific key event: 2015 Two Sessions (of the NPC and the CPPCC) in China, we would like to know the detailed public opinions towards a plethora of relatively focused themes that have aroused heated discussions, e.g., the downward pressure on GDP, the opportunities in Jing-Jin-Ji integration, the Hukou reform, anti-corruption, environment protection, *etc*. The aspect-based opinion mining technique is an intuitive candidate to fulfill this task.

The rich and varied types of online media actually mean more to us. For instance, we could expect diverse yet complementary information provided by CNN and Twitter about the 2016 Rio Olympic Games, where the former would tell more about the matches themselves and the latter instead would reflect more of public sentiment towards the matches. In other words, there is a great opportunity (or challenge from the technical side) for comprehensive public opinion analysis across different media collections. Indeed in the literature, there have been some excellent works on cross-collection topic modeling [1], [5], [6], [26]. However, they either pay less attention to the complementarity aspects across collections [5], or focus solely on topics and aspects without considering the opinions [6]. Therefore, further study is still in great need for building a cross-collection aspect-based opinion mining model, based on which the diversity and complementarity in both aspect and opinion could be learned across collections containing substantially asymmetric information, e.g., the news collection with clear aspects versus the tweet collection with strong opinions.

To address the above challenge, in this paper, we propose Cross-collection Auto-labeled MaxEnt-LDA (CAMEL), a novel topic model for complementary aspect-based opinion mining across asymmetric collections. To our best knowledge, our work is among the earliest studies in this direction. CAMEL is essentially a type of cross-collection LDA model, which models aspect-level opinions and gains information complementarity by learning both common and specific aspects across different collections. By keeping all the corresponding opinions for both common and specific aspects, CAMEL is also capable of conducting contrastive opinion analysis. Moreover, to boost CAMEL, we propose AME, an automatic labeling scheme for maximum entropy model, to discriminate aspect and opinion words without heavy human labeling. It is further enhanced to the so-called EAME scheme by employing the word embedding-based similarity. Finally, we propose CAMEL-DP, a nonparametric alternative to CAMEL. CAMEL-DP is based on coupled Dirichlet

- *Y. Zuo and J. Wu are with the School of Economics and Management, Beihang University, Beijing Shi 100191, China. E-mail: {zuoyuan, wujj}@buaa.edu.cn.*
- *H. Zhang, D. Wang, and K. Xu are with the School of Computer Science and Engineering, Beihang University, Beijing Shi 100191, China. E-mail: {hzhang, dqwang}@buaa.edu.cn, kexu@nlsde.buaa.edu.cn.*

Manuscript received 10 Oct. 2016; revised 14 Sept. 2017; accepted 9 Oct. 2017. Date of publication 18 Oct. 2017; date of current version 9 Jan. 2018. (Corresponding author: Junjie Wu.)
Recommended for acceptance by B. Poblete.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2764084

1041-4347 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

processes [16], and is capable of automatically estimating the number of common and specific aspects, which might be a headache in practice for parametric models like CAMEL.

Extensive experiments are conducted on synthetic multi-collection data sets to evaluate the performances of CAMEL and CAMEL-DP. Specifically, we design a sentence classification experiment to demonstrate that both CAMEL and CAMEL-DP can find higher-quality aspects than some state-of-the-art baseline methods. In particular, they show more robust performance in dealing with multiple collections with sample imbalance in varying degrees. Besides, Our models exhibit obvious strengths in learning more coherent opinions and more relevant aspect-opinion pairs in terms of the coherence measure. Also, the AME and EAME models indeed outperform manual labeling in distinguishing aspect words from opinion ones. Finally, a case study on a real-life public security event: 2014 Shanghai Stampede, further demonstrates the practical value of our model in real-world public opinion analysis.

The remainder of this paper is organized as follows. In Section 2, we define the complementary aspect-based opinion mining problem. In Sections 3.1 and 4, we propose the CAMEL model and give the inference method. The non-parametric alternative to CAMEL is given in Section 5. Experimental results and related works are given in Sections 6 and 7, respectively, and we finally conclude our work in Section 8.

2 Problem Definition

Our work in this paper focuses on taking advantage of aspect complementarity across multiple media collections to perform public opinion mining. Specifically, we aim to answer the following interesting questions: 1) What are the main concerns during a public event for users active on different media platforms? Do these concerns share anything in common or are they just scattered in different media? 2) What are the public opinions towards these concerns? Are they consistent or diverse in different media?

To answer these questions, we first need an *aspect-based opinion mining* model to capture public concerns and opinions simultaneously from text collections. Here "aspect" means an underlying theme, perspective or viewpoint as to a subject like an event or a product, with the assumption that each sentence is generated by a single aspect. It has been reported that about 83 percent of sentences in online reviews cover a single aspect [35], implying this assumption holds particularly for online social media such as Twitter or Chinese Weibo.

In addition, we need to build a cross-collection framework for the aspect-based opinion model so as to enable information integration from different collections. Actually a cross-collection model could benefit from this integration. Suppose we want to mine opinions from both news and micro-blogs collections. Since micro-blogs are mostly generated by public users, referred to as user generated content (UGC), we could expect sharper opinions from UGC but more vague aspects due to the more emotional and colloquial expression conventions. This is in contrast to the news, where the concerns are usually very clear but the opinions are often monotonous and implicit.

In other words, we have *asymmetric collections* or *complementary collections* for public opinion analysis. Therefore, an intuitive way to display both sides' respective advantages is to use news to help tweets identify meaningful aspects and to use tweets to enrich news through diverse opinions. This is what we called *Complementary Aspect-based Opinion Mining* and the task is defined as follows:

Given multiple text collections about a subject, design a cross-collection model, which can leverage complementary information from different collections to form aspects-based opinions for comprehensive and contrastive public opinion analysis.

There have been some studies on cross-collection topic modeling in the literature, and to our best knowledge the ones most related to our task include [5] and [6]. While [5] also studies the contrastive opinion mining problem, it does not jointly model aspects and opinions; and pays little attention to the complementarity of aspects across collections, which however is our main focus. The task of [6] is to summarize texts across complementary collections, but it ignores public opinions totally, which is also a main theme of our task.

3 CAMEL: Model and Inference

In this section, we introduce our *Cross-collection Auto-labeled Max Ent-LDA* model (*CAMEL* for short) for aspect-based opinion mining across complementary collections. Specifically, we first describe the key points of CAMEL as well as the generative process under CAMEL, which is then followed by the approximate posterior inference for CAMEL.

3.1 Model Description

As illustrated in Section 2, our main task is to design a model that can leverage complementary information from diverse collections to jointly model aspects and opinions for public opinion analysis. Along this line, two key problems need to be addressed in the model. The first one is how to model both aspects and opinions hidden in a collection simultaneously. The second problem is how to capture the complementarity across multiple collections with possible severely asymmetric information.

Let us first consider the case of aspect-based opinion mining from two asymmetric collections, e.g., the news collection with clear aspects and the tweet collection with sharp opinions. In the *aspect* level, to help extract clear aspects from tweets, it is intuitive to share with the tweet-side some similar aspects found from the news-side. This can be done by mining aspects from the two collections separately, and then linking together similar aspects from the two sides. However, it would be very difficult, if not impossible, to define a proper similarity measure and set a good threshold for it. Therefore, it would be better to design a cross-collection model that can directly mine the common aspects shared by different collections. In the *opinion* level, however, it would be more interesting to read public opinions from the tweet-side, and compare them with the opinions from the news-side, which are often regarded as the mainstream opinions from authoritative media. That means our cross-collection model should be able to mine opinions separately from different sides for the comparison purpose.

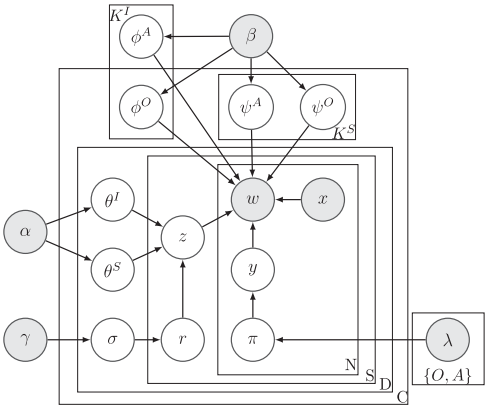


Fig. 1. Plate notation of CAMEL.

Following the above ideas, we now describe our model CAMEL, and give the generative process under it. CAMEL is essentially a cross-collection LDA model with a maximum entropy model embedded to determine the priors for aspect and opinion words switching. CAMEL assumes that different collections not only share some common aspects but also have aspects of their own. Hereinafter, we call the aspects shared across collections *common aspects*, and call the aspects only contained in one collection *specific aspects*. CAMEL also assumes that each specific aspect has a corresponding opinion. As to common aspects, CAMEL assumes each of them has C (number of collections) corresponding opinions, one for each collection, since we want to mine opinions separately from different collections for the purpose of comparison. We now describe how to generate a document under CAMEL as follows. The plate notation of CAMEL is shown in Fig. 1, with a summary of math notations given in Table 1.

Suppose there are several multinomial word distributions from a symmetric Dirichlet prior with parameter β, including: K^I common aspects $\{\phi_z^A\}_{z=1}^{K^I}$ shared by all collections, with C opinions $\{\phi_{z,c}^O\}_{c=1}^{C}$ for each ϕ_z^A, where C is the number of collections; K^S specific aspects $\{\psi_z^A\}_{z=1}^{K^S}$ and K^S corresponding opinions $\{\psi_z^O\}_{z=1}^{K^S}$, one for each collection. All these are multinomial distributions over the vocabulary containing V words in total. From Fig. 1, we see ϕ^A is placed outside the collection plate, while ϕ^O is placed inside the collection plate. This is due to the fact that common aspects are independent from collections, and the opinions of common aspects are mined separately from different collections. In contrast, ψ^A and ψ^O are both placed inside the collection plate for they are collection dependent. Note that we here assume all collections have the same number of specific aspects for simplicity, which could be relaxed to allow variation easily.

For any sentence s in document d, we draw $r_{d,s}$ from a Bernoulli distribution over $\{0, 1\}$ parameterized by σ, which in turn is drawn from a symmetric $Beta(\gamma)$. $r_{d,s}$ is an indicator of whether sentence s is generated by common or specific aspects. Specifically, when $r_{d,s} = 0$, we assume a sentence is generated by a common aspect, otherwise by a collection-specific aspect. For word n in sentence s, we introduce an indicator variable $y_{d,s,n}$ for aspect and opinion switching, which is drawn from a Bernoulli distribution over $\{0, 1\}$ parameterized by π. Similar to $r_{d,s}$, we assume a word is generated by an aspect-word distribution when $y_{d,s,n} = 0$, otherwise by an opinion-word distribution. According to some previous studies [15], [19], topic models that set π with symmetric priors are unable to identify opinion words well. Therefore, to set π for word $w_{d,s,n}$, we utilize the weights learned by the maximum entropy component and the feature vector $\mathbf{x}_{d,s,n}$ of w, which give

$$p(y_{d,s,n} = l|\mathbf{x}_{d,s,n}) = \pi_l^{d,s,n} = \frac{\exp(\lambda_l \cdot \mathbf{x}_{d,s,n})}{\Sigma_{l'=0}^{1}\exp(\lambda_{l'} \cdot \mathbf{x}_{d,s,n})},$$

where $\{\lambda_0, \lambda_1\}$ denote the weights learned by the maximum entropy model upon a set of training data, whose labels are obtained by an automatic procedure described in Section 4.

The generative process is described as follows:

1) For each common aspect z:
   a) Choose $\phi_z^A \sim Dir(\beta)$
2) For each collection c:
   a) Choose $\phi_{z,c}^O \sim Dir(\beta)$ for each common aspect z
   b) Choose $\psi_{z,c}^A, \psi_{z,c}^O \sim Dir(\beta)$ for each collection-specific aspect z
3) For each document d:
   a) Choose a collection indicator c
   b) Choose $\theta_d \sim Dir(\alpha)$
   c) Choose $\sigma_d \sim Beta(\gamma)$
   d) For each sentence s:
      i) Choose $r_{d,s} \sim Bern(\sigma_d)$
      ii) if $r_{d,s} = 0$, choose $z_{d,s} \sim Multi(\theta_d^I)$
          if $r_{d,s} = 1$, choose $z_{d,s} \sim Multi(\theta_d^S)$

TABLE 1
Math Notations

Notation	Description
K^I	the number of common aspects in total
K^S	the number of specific aspects for each collection
C	the number of collections
D	the number of documents in a collection
S	the number of sentences in a document
N	the number of words in a sentence
ϕ^A	common aspect-word distribution
ϕ^O	common opinion-word distribution
ψ^A	specific aspect-word distribution
ψ^O	specific opinion-word distribution
θ^I	common aspect mixture for a document
θ^S	collection specific aspect mixture for a document
σ	parameters for common and specific aspect switching for a sentence
π	parameters for aspect and opinion switching for a word
w	an observed word
z	aspect index for a sentence
x	feature vector for the maximum entropy (MaxEnt) model
λ	weights learned by the MaxEnt model
y	aspect and opinion switcher for a word
r	common and specific aspect switcher for a sentence
α	Dirichlet prior parameter for θ
β	Dirichlet prior parameter for all word distributions
γ	symmetric Beta prior parameters for σ
$Bern(\cdot)$	Bernoulli distribution with parameter$(\cdot)$
$Beta(\cdot)$	Beta distribution with parameter$(\cdot)$
$Multi(\cdot)$	Multinomial distribution with parameter$(\cdot)$
$Dir(\cdot)$	Dirichlet distribution with parameter$(\cdot)$

iii) For each word n:
A) Choose $y_{d,s,n} \sim Bern(\pi_{d,s,n})$
B) if $r_{d,s} = 0$ and $y_{d,s,n} = 0$, choose $w_{d,s,n} \sim Multi(\phi^A_{z_{d,s}})$
if $r_{d,s} = 0$ and $y_{d,s,n} = 1$, choose $w_{d,s,n} \sim Multi(\phi^O_{z_{d,s},c})$
if $r_{d,s} = 1$ and $y_{d,s,n} = 0$, choose $w_{d,s,n} \sim Multi(\psi^A_{z_{d,s},c})$
if $r_{d,s} = 1$ and $y_{d,s,n} = 1$, choose $w_{d,s,n} \sim Multi(\psi^O_{z_{d,s},c})$

3.2 Approximate Posterior Inference

It is obvious that exact posterior inference is intractable in CAMEL, so we turn to a collapsed Gibbs sampling algorithm [7] for approximate posterior inference, which is simple to derive, comparable in speed to other estimators, and can approximate a global maximum. Following the convention in previous work, we skip the derivation details and only present the sampling formulas. Note that the MaxEnt component (see Section 4 for more details) is trained before we perform Gibbs sampling, which means $\{\lambda_0, \lambda_1\}$ are fixed during Gibbs sampling.

In CAMEL, we have three sets of latent variables: z, r and y. Given the assignments of all other hidden variables, we can jointly sample $(z_{d,s}, r_{d,s})$ as a block

$$P(z_{d,s} = k, r_{d,s} = j | \mathbf{z}_{\neg(d,s)}, \mathbf{r}_{\neg(d,s)}, \mathbf{y}, \mathbf{w}, \mathbf{x}) \propto \frac{C^d_{(j)} + \gamma}{C^d_{(\cdot)} + 2\gamma} \times \frac{C^{d,j}_{(k)} + \alpha}{C^d_{(j)} + K^j\alpha}$$

$$\times \left(\frac{\Gamma\left(C^{A,j,k}_{(\cdot)} + V\beta\right)}{\Gamma\left(C^{A,j,k}_{(\cdot)} + N^{A,j,k}_{(\cdot)} + V\beta\right)} \cdot \prod_{v=1}^{V} \frac{\Gamma\left(C^{A,j,k}_{(v)} + N^{A,j,k}_{(v)} + \beta\right)}{\Gamma\left(C^{A,j,k}_{(v)} + \beta\right)} \right)$$

$$\times \left(\frac{\Gamma\left(C^{O,j,k}_{(\cdot)} + V\beta\right)}{\Gamma\left(C^{O,j,k}_{(\cdot)} + N^{O,j,k}_{(\cdot)} + V\beta\right)} \cdot \prod_{v=1}^{V} \frac{\Gamma\left(C^{O,j,k}_{(v)} + N^{O,j,k}_{(v)} + \beta\right)}{\Gamma\left(C^{O,j,k}_{(v)} + \beta\right)} \right).$$

We first consider the case of $j = 0$. With this condition, $C^d_{(j)}$ is the number of sentences assigned to common aspects in document d. $C^{d,j}_{(k)}$ is the number of sentences assigned to common aspect k in document d. K^j is the number of common aspects, i.e., K^I equivalently. $C^{A,j,k}_{(v)}$ is the number of times word v is assigned as an aspect word to a common aspect k, and $C^{O,j,k}_{(v)}$ is the number of times word v is assigned as an opinion word to common aspect k. $C^{A,j,k}_{(\cdot)}$ is the total number of times any word is assigned as an aspect word to common aspect k, and $C^{O,j,k}_{(\cdot)}$ is the total number of times any word is assigned as an opinion word to common aspect k. $N^{A,j,k}_{(v)}$ is the number of times word v is assigned as an aspect word to aspect k in sentence s of document d, and similarly, $N^{O,j,k}_{(v)}$ is the number of times word v is assigned as an opinion word to aspect k in sentence s of document d. When $j = 1$, all counts mentioned above refer to specific aspects. $C^d_{(\cdot)}$ is the number of sentences in document d. Note that all these counts represented by symbol C exclude sentence s of document d.

With assignments of $\mathbf{z}$ and $\mathbf{r}$, we can sample $y_{(d,s,n)}$ for $y_{(d,s,n)} = 0$

$$p(y_{(d,s,n)} = 0 | \mathbf{z}, \mathbf{r}, \mathbf{y}_{\neg(d,s,n)}, \mathbf{w}, \mathbf{x}) \propto \frac{exp(\lambda_0 \cdot \mathbf{x}_{d,s,n})}{\Sigma^1_{l'=0} exp(\lambda_{l'} \cdot \mathbf{x}_{d,s,n})} \times \frac{C^{A,r_{d,s},z_{d,s}}_{(w_{d,s,n})} + \beta}{C^{A,r_{d,s},z_{d,s}}_{(\cdot)} + V\beta},$$

and for $y_{(d,s,n)} = 1$

$$p(y_{(d,s,n)} = 1 | \mathbf{z}, \mathbf{r}, \mathbf{y}_{\neg(d,s,n)}, \mathbf{w}, \mathbf{x}) \propto \frac{exp(\lambda_1 \cdot \mathbf{x}_{d,s,n})}{\Sigma^1_{l'=0} exp(\lambda_{l'} \cdot \mathbf{x}_{d,s,n})} \times \frac{C^{O,r_{d,s},z_{d,s}}_{(w_{d,s,n})} + \beta}{C^{O,r_{d,s},z_{d,s}}_{(\cdot)} + V\beta}.$$

Here counts represented by symbol C exclude word n in sentence s of document d.

We here briefly discuss the similarities and differences between CAMEL and two most closely related models, i.e., ccTAM [6] and CPT [5], from the perspective of model structure. ccTAM distinguishes common topics from specific ones same as CAMEL did, but it does not separate opinions from topics. CPT explicitly separates opinions from topics. However, it needs strict rules to separate opinion and topic words, whereas we provide a softer way for opinion-aspect switching over a word. Besides, CPT does not distinguish common and specific topics, which limits its applicability to asymmetric collections.

4 Auto-Labeled MaxEnt Model

In this section, we illustrate how to obtain the priors for aspect and opinion switching (i.e., λ) in CAMEL. An auto-labeled maximum entropy model is proposed for this purpose, which is further enhanced by a new feature generated as the similarity of word embeddings.

4.1 AME: Auto-Labeled MaxEnt Model

In order to obtain aspect-specific opinions, we adopt the MaxEnt-LDA model proposed by Zhao et al. in [35], where a maximum entropy model (MaxEnt) is trained with Part-Of-Speech (POS) tags of words serving as priors for aspect and opinion switching. This is motivated by the fact that aspect and opinion terms normally play different syntactic roles in a sentence. But it also suffers from the high cost for manual labeling of word tags.

To address the cost issue, we propose a procedure to label training data automatically, and thus form the so-called *Auto-labeled MaxEnt model* (*AME* for short). It is motivated by the observation that opinion words usually do not appear near each other in a sentence. This, in other words, implies that a word appears next to a known opinion word is likely to be a non-opinion word. Note that this assumption is indeed based on our own observation on some product review data sets. Whether it is evident in all types of opinionated texts is still an open problem, which we would leave to the future work. The following gives the details of the procedure:

1) We first randomly select a set of M opinion words $\{V^{(O)}_m\}^M_{m=1}$ from a general opinion lexicon, which is usually publicly available for many languages. Those selected words are required to have relatively large document frequency in the target corpus.

2) We then randomly choose a set of sentences S such that each sentence in S contains at least one opinion word in $\{V_m^{(O)}\}_{m=1}^{M}$.
3) We label a word as an aspect word if it is not in $\{V_m^{(O)}\}_{m=1}^{M}$ and appears next to a known opinion word in a sentence contained by S. In this way, we finally obtain M opinion words $\{V_m^{(O)}\}_{m=1}^{M}$ and N aspect words $\{V_n^{(A)}\}_{n=1}^{N}$.

With the opinion and aspect words extracted via the above procedure, we are able to obtain POS tag features from their context to train the MaxEnt model.

4.2 EAME: Enhanced AME with Word Embedding-Based Similarity

In practice, we find the AME model with pure POS tag features works better for English rather than Chinese. Besides the dissimilarities in the languages, the major distinction stems from the POS tag tool used for preprocessing, which indicates the potential risk from using only the POS tag features. To deal with this, we here propose a new feature based on a word embedding method. Specifically, we use *word2vec*[1] to get word embeddings.

One major advantage of word2vec representation is that it learns both semantic and syntactic relations between words in an unsupervised way. After the training of word space, opinion words will closely locate in a local subspace, since they share the same syntactic role and even similar semantics. For instance, if we compute the top-five similar words of the opinion word "happy" from the reviews data used in our experiments (see Section 6 for more details), we get the following results: "satisfied", "pleased", "impressed", "satisfy" and "delighted", all of which are opinion words and have similar meanings as "happy". With the above observation, we design a new feature for the AME model based on the similarity of word embeddings and general opinion lexicons. The formula of the new feature for a target word w_t is

$$w2vFeature(w_t) = \left\lfloor \sum_{n=1}^{N} \mathrm{sgn}(w_n) \cos(w_n, w_t) \right\rfloor,$$

with

$$\mathrm{sgn}(w_n) = \begin{cases} 1 & \text{if } w_n \text{ is in the opinion lexicon} \\ -1 & \text{otherwise,} \end{cases}$$

where w_n is one of the top-N similar words of w_t, and $\cos(w_n, w_t)$ is the cosine similarity between the embeddings of w_n and w_t. To use the MaxEnt model in the subsequence procedure, we discretize the feature by rounding down its value.

Intuitively, if a word locates near to several known opinion words in the word space induced by word2vec, it is highly likely to be an opinion word. Otherwise, the word seems more likely to be an aspect word. In short, $w2vFeature(w_t)$ reflects the degree of likelihood of w_t to be an opinion word. Therefore, by using $w2vFeature(w_t)$ as supplement to POS tag features, we can alleviate the issue caused by inaccurate POS tag features, which will be testified in the experimental section below.

Here, we briefly discuss the number of training samples required by AME and EAME to train a MaxEnt model, and their dependence on the coverage of the opinion lexicon. For AME, we only take the POS tags of three words (target word and its previous and next word) as features. Since the number of categories of POS tags is only 36 in the POS tagger that we used, even dozens of training examples is able to train a MaxEnt model [35]. Therefore, dozens of sentiment words are enough for the AME model to label opinion words as well as aspect words. Which means AME does not depend on an opinion lexicon with high coverage. As to EAME, we only add one more feature based on word embedding-similarity, which can take $2N + 1$ different discrete values. Thus, few training examples are also enough for EAME to train a MaxEnt model. However, EAME relies on the opinion lexicon to obtain accurate $w2vFeature(\cdot)$, and therefore it depends on opinion lexicon coverage to some extent.

5 Nonparametric Alternative of CAMEL

Before using parametric topic models such as LDA, one has to specify the number of topics first, which in practice is either unknown or just hard to estimate, making model selection extremely hard. The situation might be even worse when applying CAMEL, since it has multiple aspects. In light of this, we provide here a nonparametric alternative of CAMEL, denoted as CAMEL-DP, based on coupled Dirichlet Processes.

5.1 Dirichlet Processes

Bayesian nonparametric models, especially those based on Dirichlet processes (DPs), have emerged as an important tool for model selection. A Dirichlet process, denoted as $DP(\alpha, B)$, is the distribution of a random probability measure D over underlying space Ω, where α is the concentration parameter and B is a base measure over Ω. With $D \sim DP(\alpha, B)$, we can draw a parameter $\theta_i \sim D$; by using it as prior for a mixture model G, we obtain the Dirichlet process mixture model (DPM). The observed data points x_i, $i = 1, \ldots, n$, are generated from $G(\theta_i)$.

One significant assumption underlying a DPM is that observations are infinitely exchangeable. This assumption does not hold in the cases when samples in different groups are not exchangeable, e.g., words (samples) from different documents (groups) are obviously not exchangeable. Hierarchical Dirichlet process (HDP) [28] is one of the most popular approaches to address this issue, which organizes DPs into a tree structure with the parents serving as base measures for the children. A two-level HDP can ensure that the sets of group-specific DPs share the same atoms. Specifically, corpus-level DPs provide base measures for document-level DPs, thus all documents share a same set of atoms (topics).

When there are multiple corpora, a three-level HDP is more appropriate. In detail, a top-level DP yields the base measure for each of the corpus-level DPs. Draws from each of these corpus-level DPs yield the base measures for DPs associated with the documents within a corpus. This three-level HDP outperforms the two-level one, since the latter ignores the document-corpus membership.

1. https://code.google.com/p/word2vec/

Nevertheless, the three-level HDP does not distinguish common and specific topics, which is indeed very important to our complementary aspect-based opinion mining. Therefore, we adopt coupled Dirichlet processes to build the nonparametric version of CAMEL.

5.2 CAMEL-DP

In order to model common and specific aspects, we introduce two kinds of DPs. One is the global DP shared by different collections, and the others are local DPs, one for each specific collection. To generate the ith observation (sentence) s_{ji} of the jth group (document) d_j in collection c, we first draw the parameter θ_{ji} of the mixture model G from F_c, and then generate observations as $s_{ji} \sim G(\theta_{ji})$.

The significant part is F_c, which is a combination of two components, D_0 and D_c. D_0 is a global component shared by all collections, which is drawn from a global DP. D_c is a specific component of collection c, which is drawn from a local DP. Formally, the generative process of F_c is

$$\begin{aligned} &D_0 \sim DP(\alpha_0, B_0),\\ &D_c \sim DP(\alpha_c, B_c),\\ &\epsilon_c \sim Beta(\alpha_0, \alpha_c),\\ &F_c = \epsilon_c D_0 + (1-\epsilon_c) D_c. \end{aligned}$$

Due to the nature of F_c, the mixture model G of document j inherits common aspects from the global DP as well as specific ones from the local DP of collection c. As a result, CAMEL-DP can simultaneously model common and specific aspects, just like the parametric CAMEL, for documents coming from multiple collections.

Before going into the details of the inference of CAMEL-DP, it is noteworthy that *F_c is still a sample from a DP*. This guarantees that the mixture model G can share aspects generated from D_0 and D_c, and simplifies the inference of CAMEL-DP. According to Lin et al. [16], one can construct new DPs by conducting three operations on existing ones. F_c is constructed by performing *superposition* of global and local DPs, which is one of the three operations. Readers with more interest along this line can refer to Lin's paper for the details of the three operations.

5.3 Inference of CAMEL-DP

This section introduces a Gibbs sampling algorithm to estimate CAMEL-DP. We first set up the notations. Assume there are J groups of data, one global DP and C local ones, with C being the number of collections. The observations in the jth group are $s_{j1}, \ldots, s_{jn_j}$. An atom is denoted as ϕ_k, where k is a globally unique identifier of the atom. Instead of instantiating θ_{ji} for each data sample s_{ji}, we assign an indicator z_{ji} to it, which is equivalent to setting $\theta_{ji} = \phi_{z_{ji}}$. To facilitate the sampling process, for each atom ϕ_k, we maintain an indicator e_k specifying which DP contains it, the global one or a specific local one, and a set of counters $\{m_{jk}\}$, where m_{jk} stores the number of data samples associated with atom k in group j. We also maintain a set I_u for D_u (the uth DP), which contains the indices of all atoms in D_u.

Each data sample s_{ji} is assigned with a latent label z_{ji}. To draw z_{ji}, we first have to choose the global DP or the local one as the source. We use r_{ji} to denote the source DP of z_{ji}. Specifically, $r_{ji} = 0$ indicates the global DP is the source, $r_{ji} = c$ indicates the local DP of collection c (where the jth group belongs to) is the source. The sample equation for r_{ji} is

$$\begin{aligned} p(r_{ji} = u) &\propto p(r_{ji} = u|\boldsymbol{v}_j) p(s_{ji}|r_{ji} = u, \boldsymbol{z}_{\neg ji})\\ &= v_{ju} p(s_{ji}|r_{ji} = u, \boldsymbol{z}_{\neg ji}), \end{aligned}$$

where $\boldsymbol{v}_j = (v_{j0}, v_{jc})$ are the group-specific priors over DP sources, *a.k.a.* the combination coefficients. $p(s_{ji}|r_{ji} = u, \boldsymbol{z}_{\neg ji})$ is the likelihood of s_{ji}

$$\begin{aligned} &p(s_{ji}|r_{ji} = u, \boldsymbol{z}_{\neg ji})\\ &= \frac{1}{w_{uj\neg i} + \alpha_u} \left(\sum_{k \in I_u} m_{*k_{\neg ji}} f(s_{ji}; \boldsymbol{\phi}_k) + \alpha_u f(s_{ji}; B) \right), \end{aligned}$$

where $m_{*k_{\neg ji}}$ is the total number of samples assigned to k in all groups except for s_{ji}, $w_{uj\neg i} = \sum_{k \in I_u} m_{*k_{\neg ji}}$, $f(s_{ji}; \boldsymbol{\phi}_k)$ is the pdf at s_{ji} *w.r.t.* $\boldsymbol{\phi}_k$, and $f(s_{ji}; B) = \int_\theta f(s_{ji}; \theta) B(\theta) d\theta$.

Once a DP is chosen, we can draw a particular atom. The process is similar to the Chinese restaurant process: with a probability proportional to $m_{*k_{\neg ji}} f(s_{ji}; \boldsymbol{\phi}_k)$, we set $z_{ji} = k$, and with a probability proportional to $\alpha_u f(s_{ji}; B)$, we draw a new atom from $B(\cdot|s_i)$.

Along with updating $\boldsymbol{z}$, we also have to update the combination coefficients. The coefficient $\boldsymbol{v}_j = (v_{j0}, v_{jc})$ reflects the relative contribution of the global DP or local one to the jth group. $\boldsymbol{v}_j$ follows a Beta distribution, according to the generative process of F_c. Given $\boldsymbol{z_j}$, we have

$$(v_{j0}, v_{jc}|\boldsymbol{z_j}) \propto Beta(\alpha_0 + \sum_{k \in I_0} m_{jk}, \alpha_c + \sum_{k \in I_c} m_{jk}).$$

Here $\sum_{k \in I_c} m_{jk}$ is the total number of samples in the jth group that associate with D_c.

Note that the atom $\boldsymbol{\phi}_k$ is actually a pair of multinomial distributions in our application, one for the aspect-word distribution, and the other for the opinion-word distributions. Besides, the sampling of y (aspect-opinion word switcher) in CAMEL-DP is the same as in CAMEL.

6 Experimental Results

In this section, we present extensive experimental results to evaluate CAMEL and CAMEL-DP. Hereinafter, we agree to use "CAMEL" and "ours", "CAMEL-DP" and "oursNP" interchangeably in comparative studies. We also use "our methods" to denote both "CAMEL" and "CAMEL-DP" occasionally for concision.

6.1 Experimental Setup

6.1.1 Data Sets

Our methods are tested on two real-world data sets. One is a text collection of online reviews for electronic devices on Amazon,[2] which is reorganized so as to provide a control for evaluation. The other is a text collection about a real-world public event crawled from news portals as well as Chinese Weibo,[3] to further verify the practical use of our methods.

2. http://www.amazon.com
3. http://www.weibo.com

TABLE 2
Statistical Description of Data Sets

Data	#Documents	#Sentences	#Words
*Amazon_*C_0	3,535	56,053	315,471
*Amazon_*C_1	3,659	60,935	339,192
Stampede_News	1,015	42,651	262,613
Stampede_Tweets	13,004	25,425	74,353

The online reviews were originally collected by Jo et al. [13], which contains reviews for electronic devices in seven categories. To evaluate the quality of common and specific aspects as well as their associated opinions across collections, we create a new data set based on those reviews. Specifically, we first select reviews under three categories, namely *coffee machine*, *canister vacuum* and *MP3 player*, each of which shares minimal overlap in contents with others. We then place reviews labeled as "canister vacuum" into one collection and reviews labeled as "MP3 player" into the other collection, and randomly inject sentences of "coffee machine" reviews into these two collections, which forms the two data sets *Amazon_*C_0 (C_0 for short) and *Amazon_*C_1 (C_1 for short), respectively, in Table 2. In this way, the reviews for "coffee machine" are fragmented and scattered over C_0 and C_1, which artificially constructs the complementarity between the two collections when we want to recover the review aspects for "coffee machine". In other words, we could expect common aspects about "coffee machine" across the two collections, and specific aspects about "canister vacuum" and "MP3 Player" in C_0 and C_1, respectively. As to the real-word event data set, we crawled news and tweets related to event *2014 Shanghai stampede*, which is a sad public accident with around 40 deaths due to an overcrowding stampede in Chenyi Square, Shanghai, Dec. 31, 2014.

All data sets had gone through the same preprocessing procedure as follows. We first apply POS tagging, and then get automatically labeled training data from our AME or EAME model, as described in Section 4. We finally remove stop words and those with low document frequencies, and also remove URLs and hashtags for tweets. Note that, the word embeddings used in EAME are trained on datasets used in this paper, i.e., we do not apply any pre-trained word embeddings. We use Stanford POS Tagger[4] to tag English online reviews and LTP-Cloud[5] to tag Chinese news and tweets. The opinion lexicon used for English corpora is collected by Hu and Liu [10]. For Chinese corpora we use an opinion lexicon merged from two widely used Chinese opinion lexicons, i.e., the NTU Sentiment Dictionary (NTUSD)[6] and the Chinese/English Vocabulary for Sentiment Analysis (VSA).[7] Details of the preprocessed data sets are shown in Table 2.

6.1.2 Baseline Methods

In the experiments, we compare our methods with two baseline methods namely LocLDA [4] and MaxEnt-LDA [35]. The ccTAM [6] can be a candidate baseline if we allow for ambiguity on the "topic" and "aspect" concepts. Therefore, we also conduct comparison with ccTAM but leave details to the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2017.2764084. In order to learn aspects instead of topics, LocLDA performs standard LDA on a collections of sentences (it treats each sentence as a document). MaxEnt-LDA assumes each sentence is generated from a single topic, and therefore topics induced by it also map to aspects. MaxEnt-LDA further separates aspect and opinion words such that a topic is factorized into two dimensions, one for aspects and the other for opinions. For concise expression, we refer LocLDA as BL0 and MaxEnt-LDA as BL1. We perform the two baseline methods over different Amazon collections and summarize the notations for baselines in Table 3. For instance, LocLDA over collection *Amazon_*C_0 is referred to as BL0-0, and so forth. Note that CAMEL and CAMEL-DP are designed purposefully for multiple collections, so it is unnecessary to add "-2" as the suffix to them. Comparison of our methods with these baselines over different collections can give us insights about whether aspects and opinions induced by our methods can benefit from the complementarity between collections. We will give evidence in the experiments to follow.

TABLE 3
Baseline Notations

Notation	Description
C_0	the *Amazon_*C_0 collection
C_1	the *Amazon_*C_1 collection
BL0	LocLDA [3]
BL1	MaxEnt-LDA [35]
BLi-0	Performing baseline i over C_0
BLi-1	Performing baseline i over C_1
BLi-2	Performing baseline i over C_0&C_1

6.1.3 Parameter Setting

In our practice, topic models with weak priors often perform better on short texts. Therefore, we set $\alpha = 0.1$ and $\beta = 0.01$ for MaxEnt-LDA, LocLDA and CAMEL. We set $\gamma = 0.1$ for both CAMEL and CAMEL-DP. For CAMEL-DP, we set concentration parameter $\alpha = 0.2$ for each DP. The base distribution B of CAMEL-DP is assumed to be $Dir(0.05)$. The parameter setting of CAMEL-DP is to ensure it learns a similar number of aspects as CAMEL does. Gibbs sampling is employed for model inference, with 1,000 iterations for all parametric methods and 2,000 iterations for CAMEL-DP. Each parameter configuration runs 10 samples to obtain averaged results.

The number of aspects learned by CAMEL is different from the number of opinions. That is, if CAMEL learns $K^I + CK^S$ aspects, then it learns $C(K^I + K^S)$ opinions, where K^I (or K^S) is the number of common (or specific) aspects and C is the number of collections. Hence, for a fair comparison with baseline methods, the strategy used to set the number of aspects in aspect evaluation is different from the one used to set the number of opinions in opinion evaluation. Please refer to Sections 6.3 and 6.4 for detailed settings.

4. http://nlp.stanford.edu/software/tagger.shtml
5. http://www.ltp-cloud.com/
6. http://nlg18.csie.ntu.edu.tw:8080/lwku/pub1.html
7. http://www.keenage.com/html/e_index.html

TABLE 4
Comparison Among MME, AME, and EAME

Size	$P@5$			$P@10$			$P@20$		
	MME	AME	EAME	MME	AME	EAME	MME	AME	EAME
$S=10$	**0.90**	0.64	0.86	**0.82**	0.58	**0.82**	0.71	0.51	**0.77**
$S=20$	0.80	0.64	**0.92**	0.74	0.56	**0.87**	0.67	0.50	**0.81**
$S=30$	0.86	0.84	**0.91**	0.82	0.81	**0.89**	0.70	0.76	**0.83**
$S=40$	0.80	0.84	**0.93**	0.75	0.85	**0.90**	0.71	0.78	**0.84**
$S=50$	0.82	0.90	**0.96**	0.81	0.83	**0.93**	0.71	0.76	**0.87**

6.1.4 Evaluation Measures

Two sets of measures are to be used in our experiments, one is *macro-averaged precision* (p), *recall* (r) and *f-measure* (f) for aspect evaluation, and the other is *coherence* to evaluate opinions as well as aspect-opinion associations.

Since each online review belongs to one product category, we can leverage this information to evaluate the quality of the learned aspects externally. In detail, we first manually label each aspect with one of the three product categories. Since all methods assign one aspect to each sentence, this aspect label is then regarded as the sentence label given by the methods. On the other hand, each sentence is also labeled by the category where the review resides in, which can serve as the control. As a result, we can perform sentence classification to evaluate the aspects indirectly via p, r and f. Section 6.3 below gives more details. Note that we also designed an automatic procedure to label each aspect with one of the three product categories, which performs similarly to the manual one. Readers with interests can find it in the supplementary material, available online.

To compare the quality of opinions, we choose an automatic measure called *topic coherence* [20], which has been widely used in evaluating topics and has been justified in according with human evaluations. We further slightly modify the coherence score to measure the relevance between aspect and its opinions. Details can be found in Section 6.4

6.2 Validity of Auto-Labeled MaxEnt

Before giving the details of the aspects and opinions evaluations, we first justify the validity of our Auto-labeled MaxEnt (AME for short) component. We also have interests in whether new features via word embedding-based similarity (as suggested in Section 4.2) can enhance the performance of AME, and thus include EAME for comparison. The manually labeled MaxEnt (MME for short) method is adopted here as the baseline.

We compare the $Precision@n$ ($P@n$ for short) of the above models on reviews data with varying number of training sentences S. Here $P@n$ measures how many words are precisely opinion rather than aspect words given the probabilistic top-n words of an opinion, which is then validated by human. For MME, we randomly select S sentences with opinion words and manually label them. As to AME and EAME, we use our procedure to acquire the same number of training sentences and label them automatically. We increase the size of training data, and compare the $P@5$, $P@10$ and $P@20$ values of various methods as reported in Table 4.

From the results in Table 4, we observe that AME is less accurate than MME as the training size is less than or equal to 30. As the training size gets larger, however, AME outperforms MME gradually. This demonstrates that AME is not only an efficient but also an effective method for the MaxEnt component in our model, especially when the training data is of a large scale. Another observation is the surprise from EAME. That is, EAME seems comparable to MME even the training size gets as small as 10; it then continuously outperforms MME and AME as the training size gets larger. This implies that the new features by word embedding based similarity (i.e., word2vec) indeed improve AME significantly. Therefore, in the remaining experiments, we agree to use EAME as the default MaxEnt component for MaxEnt-LDA, CAMEL and CAMEL-DP.

6.3 Aspect Evaluation

We design a sentence classification experiment on review data to evaluate the quality of aspects. MaxEnt-LDA and our methods assign one aspect to each sentence, and each aspect is manually labeled by one category. As a result, we can use sentence classification as indirect evaluation for aspects—better classification results indicate higher quality of aspects. As to LocLDA, since it assigns one aspect to each word, we use the following equation to infer the aspect of a sentence

$$k = \arg\max_{k \in S^A} \sum_{n=1}^{N_{d,s}} \log P(w_{d,s,n}|k),$$

where S^A is the inferred aspect set. The two Amazon review data sets listed in Table 2 are used for this experiment, with data generation details described in Section 6.1.1. It is worth noting that the sentences from the product category "coffee machine" are scattered over the two data sets, which creates the complementarity between the two collections while learning the aspects about "coffee machine".

Procedure. We run baselines and our methods over the data to get inferred aspect set S^A and aspect assignment k of each sentence. To perform sentence classification evaluation, we manually map each aspect to one of the three categories via a mapping function $f(k) : S^A \rightarrow L$. Aspects cannot be mapped to any category are labeled as *other*. Given the mapping results, we can get the predicted labels of sentences for each method. We then evaluate all methods according to the external metrics: *precision* (p), *recall* (r) and *f-measure* (f).

Settings. We set the aspect number for BL0-2 and BL1-2 to K, and set common-aspect number to K^I and specific-aspect number to K^S in CAMEL. To keep all methods comparable, we let $K^I + CK^S = K$, where $C = 2$ is the number of collections. We vary K from 10 to 40, and set $K^I = 2, 4, 6, 8$ and $K^S = 4, 8, 12, 16$ for $K = 10, 20, 30, 40$, respectively. Here we set $K^S = 2K^I$ because we know the approximate proportion of common aspects in each collection. When dealing with data of no prior knowledge, however, one could try several configurations of K^I and K^S to find the appropriate setting, or resort to nonparametric Bayesian inference, such as CAMEL-DP described in Section 5.

Results. We report sentence classification results of all classes over different collections in Table 5 and those of the common class in Table 6. Results of specific classes are

TABLE 5
Sentence Classification Results of All Three Categories

Data	Method	K=10			K=20			K=30			K=40		
		p	r	f	p	r	f	p	r	f	p	r	f
C_0	BL0-2	**0.943**	0.675	0.783	**0.946**	0.649	0.769	**0.957**	0.595	0.733	**0.964**	0.585	0.727
	BL1-2	0.845	0.799	0.818	0.874	0.790	0.829	0.877	0.774	0.822	0.877	0.760	0.814
	CAMEL	0.848	**0.845**	**0.844**	0.876	**0.828**	**0.851**	0.877	**0.813**	**0.843**	0.866	**0.829**	**0.847**
C_1	BL0-2	**0.958**	0.698	0.804	**0.958**	0.668	0.786	**0.971**	0.635	0.767	**0.973**	0.598	0.740
	BL1-2	0.899	0.802	0.848	0.901	0.803	0.848	0.897	0.797	0.844	0.902	0.787	0.840
	CAMEL	0.861	**0.862**	**0.858**	0.896	**0.833**	**0.863**	0.897	**0.832**	**0.862**	0.887	**0.843**	**0.864**
C_0&C_1	BL0-2	**0.918**	0.683	0.779	**0.915**	0.661	0.766	**0.931**	0.617	0.741	**0.940**	0.599	0.730
	BL1-2	0.849	0.808	0.827	0.845	0.816	0.828	0.852	0.803	0.825	0.858	0.787	0.820
	CAMEL	0.870	**0.865**	**0.865**	0.891	**0.844**	**0.866**	0.896	**0.836**	**0.865**	0.885	**0.846**	**0.865**

"p", "r", and "f" stands for precision, recall, and f-measure, respectively.

TABLE 6
Sentence Classification Results of the "Coffee Machine" Category

Data	Method	K=10			K=20			K=30			K=40		
		p	r	f	p	r	f	p	r	f	p	r	f
C_0	BL0-2	**0.916**	0.698	0.791	**0.929**	0.650	0.764	**0.936**	0.606	0.735	**0.944**	0.569	0.709
	BL1-2	0.790	0.777	0.779	0.845	0.735	0.785	0.840	0.731	0.782	0.832	0.728	0.776
	CAMEL	0.805	**0.803**	**0.801**	0.851	**0.777**	**0.812**	0.846	**0.768**	**0.804**	0.827	**0.791**	**0.808**
C_1	BL0-2	**0.940**	0.700	0.800	**0.943**	0.651	0.770	**0.970**	0.607	0.746	**0.971**	0.568	0.716
	BL1-2	0.860	0.776	**0.816**	0.877	0.742	0.803	0.865	0.739	0.796	0.868	0.734	0.795
	CAMEL	0.808	**0.832**	0.814	0.862	**0.803**	**0.830**	0.864	**0.793**	**0.826**	0.843	**0.819**	**0.831**
C_0&C_1	BL0-2	**0.928**	0.699	0.795	**0.936**	0.650	0.767	**0.953**	0.607	0.741	**0.957**	0.568	0.713
	BL1-2	0.822	0.777	0.797	0.861	0.739	0.794	0.853	0.735	0.789	0.850	0.731	0.785
	CAMEL	0.806	**0.818**	**0.808**	0.857	**0.790**	**0.821**	0.855	**0.781**	**0.815**	0.835	**0.805**	**0.820**

"p", "r", and "f" stands for precision, recall, and f-measure, respectively.

omitted here to avoid redundancy, which in fact are similar to the results in the above tables. Since CAMEL-DP learns the number of aspects automatically, we present its result alone in Table 7. Note that the column "Method" indicates the method and data set that generate the aspects, and the column "Data" indicates the data set for classification evaluation. For instance, the second line of Table 5 gives the classification results over the sentences of C_0, based on the aspects learned by LocLDA over C_0&C_1.

From the comparative results in both Tables 5 and 6, it is obvious that CAMEL consistently outperforms the two baseline methods in terms of *r* and *f*. This well demonstrates the advantages of CAMEL in modeling both common and specific aspects hidden inside C_0 and C_1. To better understand this, it is worth noting that BL0-2 and BL1-2 perform aspect-opinion mining directly on combined collections without modeling the common aspects explicitly, and hence have difficulties in uncovering the fragmented reviews for "coffee machine". In contrast, CAMEL explicitly separates aspects shared by collections and those specific to each collection. As a result, common aspects in one collection can serve as the complementary information for aspects extraction in other collections in a mutual way, which finally leads to the better performances.

It is also intriguing the way CAMEL always shows lower precision values than the baselines. To this end, we traced the misclassified sentences and found that a majority of them correspond to an interesting common aspect learned by CAMEL, which we term as *after-sales service*. Table 8 gives an example of the *after-sales service* aspect-opinion words, where "opinion$i$" means the opinion induced from C_i, $i = 0, 1$. As can be seen from the Table, this aspect is mainly about the repair or replacement issues of defective products after sale, which is undoubtedly a common aspect of the reviews for all the three Amazon product categories—although we had only set up the common aspect from the "coffee machine" reviews, as designed in Section 6.1.1. From this viewpoint, CAMEL is indeed capable of extracting common aspects across different collections. But on the other hand, since this *unexpected* common aspect comes from the review sentences of all products, it essentially does harm to the classification accuracy, and results in lower precision values for CAMEL over time.

We finally turn to the performance of CAMEL-DP. As can be seen from Table 7, while CAMEL-DP performs slightly worse than CAMEL (Table 5), it outperforms BL0 and BL1 significantly on all three Amazon collections in terms of f. This illustrates that CAMEL-DP can take advantage of complementary aspects across asymmetric collections as CAMEL does, although it is less effective

TABLE 7
Sentence Classification Results of CAMEL-DP

Data	Precision	Recall	F-measure
C_0	0.867	0.805	0.834
C_1	0.886	**0.829**	**0.856**
C_0&C_1	**0.889**	0.826	**0.856**

TABLE 8
"After-Sales Service" Aspect and Opinions

aspect	service call customer back return product problem amazon buy warranty
opinion0	send free good great ship local wrong long happy fast
opinion1	send work great free good local defective creative original easy

than CAMEL in aspect learning. A difference between the inference procedures of CAMEL and CAMEL-DP is that a sentence's aspect and its common or specific aspect indicator variable are sampled as a block in CAMEL while separately in CAMEL-DP. Such a difference might cause CAMEL-DP to be less reliable than CAMEL in assigning common or specific aspects to a sentence, which ultimately degrades the sentence classification performance of CAMEL-DP as compared to CAMEL.

6.4 Opinion Evaluation

Here, we give evaluations of opinions learned by our methods. We employ *topic coherence* for opinion evaluation, upon which a new measure is further proposed to evaluate the relevance between an aspect and its corresponding opinions.

6.4.1 Opinion Coherence

Coherence score measures a single word distribution by computing the semantic similarity degree between high probability words in it. A higher score often indicates better quality. Given T high-probability words of an opinion, the coherence score for the opinion is defined as

$$C_O(k; V^{(k)}) = \sum_{t=2}^{T} \sum_{l=1}^{t-1} \log \frac{D(v_t^{(k)}, v_l^{(k)}) + \epsilon}{D(v_l^{(k)})},$$

where $V^{(k)} = (v_1^{(k)}, \ldots, v_T^{(k)})$ is a list of T the most probable words in opinion k. $D(v)$ counts the number of documents containing the word v, and $D(v, v')$ counts the number of documents containing both v and v'. ϵ is a smoothing variable used to avoid taking the log of zero for words that never co-occur.

6.4.2 Aspect-Opinion Coherence

While the coherence measure mentioned above can evaluate the quality of an opinion, it can not evaluate the relevance (i.e., coherence) between one aspect and its opinions. To meet this need, we here propose a new measure for the evaluation of aspect-opinion pairs. Given T high probability words of an aspect and its opinion, respectively, the coherence of the aspect-opinion pair is defined as

$$C_{A,O}(k; V^{A,(k)}, V^{O,(k)}) = \sum_{t=1}^{T} \sum_{l=1}^{T} \log \frac{D(v_t^{A,(k)}, v_l^{O,(k)}) + \epsilon}{D(v_t^{A,(k)})},$$

where $V^{A,(k)}$ is a list of T the most probable words in aspect k, and $V^{O,(k)}$ is a list of T the most probable words in opinion k. Loosely speaking, the value of $\frac{D(v^{A,(k)}, v^{O,(k)})}{D(v^{A,(k)})}$ estimates the probability one could observe the opinion word $v^{O,(t)}$ if he (or she) has already observed the aspect word $v^{A,(t)}$ in a document.

TABLE 9
Results of Opinion and Aspect-Opinion Coherence

Data	Method	opinion coherence		aspect-opinion coherence	
		$T = 10$	$T = 15$	$T = 10$	$T = 15$
C_0	BL1-0	-121.1 ± 3.6	-306.0 ± 6.3	-229.3 ± 5.3	-571.8 ± 7.3
	CAMEL	-119.1 ± 3.9	-306.1 ± 7.2	-225.7 ± 5.4	-569.5 ± 12.2
	CAMEL-DP	$\mathbf{-86.3} \pm 22.8$	$\mathbf{-235.5} \pm 50.1$	$\mathbf{-156.4} \pm 45.3$	$\mathbf{-420.4} \pm 101.7$
C_1	BL1-1	-129.6 ± 3.8	-332.1 ± 9.5	-246.7 ± 5.5	-621.0 ± 11.6
	CAMEL	-127.7 ± 1.7	-325.5 ± 4.4	-245.1 ± 3.2	-613.1 ± 4.8
	CAMEL-DP	$\mathbf{-89.8} \pm 18.7$	$\mathbf{-238.5} \pm 43.1$	$\mathbf{-161.1} \pm 42.1$	$\mathbf{-424.6} \pm 95.1$
C_0&C_1	BL1-2	-134.7 ± 2.5	-344.1 ± 6.8	-255.1 ± 3.5	-639.1 ± 9.1
	CAMEL	-129.7 ± 2.0	-329.1 ± 3.9	-240.5 ± 3.4	-604.0 ± 6.4
	CAMEL-DP	$\mathbf{-90.6} \pm 20.4$	$\mathbf{-242.8} \pm 45.0$	$\mathbf{-161.2} \pm 42.2$	$\mathbf{-427.6} \pm 95.2$

For both coherence scores, we set $\epsilon = 10^{-12}$ to reduce the score for completely unrelated words, as suggested in [27].

6.4.3 Opinion Evaluation Results

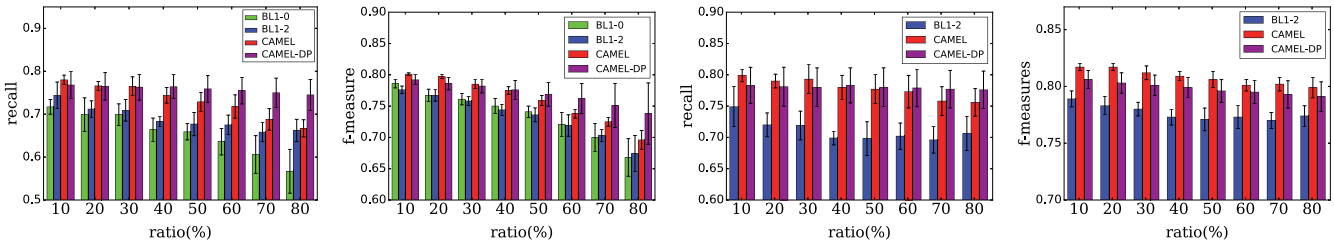
We compare our methods with MaxEnt-LDA (i.e., BL1) in terms of averaged opinion coherence and aspect-opinion coherence. Since LocLDA (i.e., BL0) does not learn opinions, we leave it out of this account. To make all methods comparable, for CAMEL with C different collections, we set $C(K^I + K^S) = K$. We set $K_0 = K_1 = 15$, $K = 30$, and $K^I = 5, K^S = 10$, where K_0 (or K_1) is the aspect number for BL0-0 (or BL0-1) and BL1-0 (or BL1-1). Note that the number of opinions for BL0 and BL1 equals to the number of aspects. Before computing the coherence score for CAMEL-DP, we remove aspect-opinion pairs assigned with less than 200 sentences. This is due to the fact that CAMEL-DP is a nonparametric model, which tends to learn new even long-tail aspects. Therefore, the aspects assigned with a handful sentences are suspicious and should be removed before evaluation.

The average results of opinion coherence and aspect-opinion coherence with $T = 10$ and $T = 15$ are listed in Table 9. As can be seen from the table, no matter which data set is used, CAMEL outperforms MaxEnt-LDA in terms of both opinion coherence and aspect-opinion coherence with varying T. This well demonstrates the advantages of CAMEL in leveraging the complementarity between collections; that is, it not only helps to find more accuracy aspects but also improves the opinion quality and the quality of aspect-opinion pairs.

It is also very interesting that while CAMEL-DP achieves the highest average coherence scores, it also suffers from the highest volatility in those scores. To understand this, recall that we run CAMEL-DP ten times to report the above results. In some runs, we occasionally find that one or two aspects learned by CAMEL-DP contain only hundreds of sentences, and thus lead to a relatively high coherence score. Note that while less supported by training data, these aspects can be found due to the usage of nonparametric priors. This well illustrates why CAMEL-DP obtains high coherence scores but with high volatility.

6.5 Value of Complementarity

In this section, we evaluate our methods with varying levels of complementarity across collections. This can give us insights about the real value of learning from complementary



(a) C_0: recall. (b) C_0: f-measure. (c) $C_0\&C_1$: recall. (d) $C_0\&C_1$: f-measure.

Fig. 2. Sentences classification results of "coffee machine" category with varying removing ratio.

information. To this end, we reform the online reviews data in Table 2 by removing the sentences for "coffee machine" from collection C_0 while keeping collection C_1 unchanged. We remove the sentences gradually from $r = 10$ to 80 percent of the total "coffee machine" sentences in C_0, and observe the changes of sentence classification results on C_0 and $C_0\&C_1$. MaxEnt-LDA is adopted here as the baseline method. We set $K_0 = K_1 = 15$, $K = 30$, $K^I = 6$ and $K^S = 12$. The resulting *recall* and *f-measure* values are illustrated in Fig. 2.

From Figs. 2a and 2b, we see that with the removal of "coffee machine" sentences, the recall and f-measure of BL1-0 drop rapidly and become more and more unstable. Since BL1-2 and our methods all can utilize the complementary information about "coffee machine" sentences in C_1, they all outperform BL1-0 significantly and seem relatively more stable. This well demonstrates the great value of learning from complementary information. Another important observation from Fig. 2 is that BL1-2 is generally less effective than our methods, and becomes rather unstable in terms of f when r reaches 80 percent. This is due to the fact that while BL1-2 uses "coffee machine" sentences from both collections, it does not model the complementarity explicitly via common aspects. As a result, the fragmented information about "coffee machine" cannot be well recovered. This observation reveals that explicit separation of common and specific aspects in CAMEL is critically important to complementary aspect mining. Finally, to our surprise, CAMEL-DP performs fairly well against other methods on collection C_0, although the volatilities of performances are still a bit higher. We believe the ability of nonparametric models in learning long-tail aspects contributes to the above interesting results. More specifically, a nonparametric model does not constrain the number of aspects, and therefore tends to extract aspects supported by less training data (i.e., aspects in the tail). In Figs. 2c and 2d, CAMEL-DP seems merely comparable to or less effective than CAMEL, but it yet outperforms BL1-2 significantly.

In short, our methods can model complementary information hidden inside different collections explicitly via the introduction of common aspects, which is very important to the aspect learning from multiple collections.

6.6 Case Study

Here we use a real-life case: *2014 Shanghai Stampede* (hereinafter referred to as *stampede* for short) to validate the effectiveness of CAMEL in practice. This is about the deadly stampede occurred on December 31, 2014 in Shanghai, near the Chen Yi Square on the Bund, where about 300,000 people had gathered for the new year celebration. In total 36 people were killed and 49 injured in this tragic event. We collected news and tweets related to the stampede as two complementary data sets (see Table 2 for details) and apply CAMEL to them with the parameter settings as follows: $\alpha = 0.1$, $\beta = 0.01$, $K^I = 10$, $K^S = 15$, where K^I and K^S refer to the numbers of common and specific aspects, respectively.

Let us first observe the common aspects learned by CAMEL. Fig. 3 exhibits three sample common aspects with clear semantic meanings. Specifically, common aspect 1 is about the so-called *fallback shouters*, which refers to a bunch of people who yelled "fallback" to alert the crowd the

Common Aspect 1		**Common Aspect 2**		**Common Aspect 3**	
人群 现场 外滩 警察 后退哥 新年 事件 游客 秩序 声音		遇难者 名单 事件 外滩 上海 身份 踩踏 广场 上午 年龄		活动 管理 措施 预案 责任 事故 政府 群众 人流 公共	
crowd scene bund policeman fallback_shouters new_year event tourist order voice		victims namelist event bund Shanghai identity stampede square forenoon age		event management precaution emergency_plan responsibility accident government crowd people public	
opinion 0	**opinion 1**	**opinion 0**	**opinion 1**	**opinion 0**	**opinion 1**
后退(recede)	上面(above)	默哀(grieve)	公布(announce)	安全(secure)	安全(secure)
好样的(great)	摔倒(tumble)	同情(sympathize)	核实(verify)	负责(responsible)	落实(implement)
感动(touched)	疑似(suspected)	祈福(blessing)	受伤(injure)	重大(major)	检查(inspect)
冷静(calm)	根本(at all)	痛惜(regret)	确认(check)	防范(prevent)	全面(overall)
希望(hope)	下来(down)	数衍(perfunctory)	拥挤(crowed)	追究(investigate)	控制(control)
脆弱(tender)	周围(around)	无知(ignorant)	发生(happen)	限制(limit)	重要(important)
减少(reduce)	拥挤(crowded)	心痛(distressed)	慰问(sympathise)	重视(take_seriously)	及时(in_time)
关键(important)	稳定(stable)	安抚(appease)	哀悼(grieve)	控制(control)	防范(prevent)
危险(dangerous)	吵杂(noisy)	简单(simple)	深切(heartfelt)	注意(be_careful)	严格(strict)
感谢(thank)	附近(nearby)	庆幸(fortunately)	悼念(mourn)	学习(learn)	完善(complete)

Fig. 3. Sample common aspects and opinions induced from the stampede data.

Aspect:	活动 管理 措施 预案 责任 事故 政府 群众 人流 (event management precaution emergency_plan responsibility accident government crowd people public)
Tweets:	1. 流量已经超出许多，早应该禁行。 Should forbid people from entering earlier, since pedestrian volume has exceeded the limitation. 2. 以后如果遇到踩踏事故又不幸跌倒，立刻双手抱头，蜷缩一团，护住身体重要部位，可以增加存活率。 If stampede accident happens to you and you unfortunately fail down, it is important to keep your hands up around your head and to crouch your body to protect yourself. 3. 北京地铁站也是很挤的，悲痛这个事情的同时在那边工作的朋友一定要注意安全! The subway stations in Beijing are also very crowded, so friends who work there should pay attention to their own safety while sorrowing about the accident! 4. 强烈建议公共场所的台阶改成缓坡，以减少绊倒的机会。 Strongly suggest changing the stairs with more gentle slopes so as to decrease the chance to fall. 5. 一路走好～愿你们在天堂得到幸福，希望监管部门以后能控制好这种人多的局面。 Farewell～wish you happy in heaven, and hope the supervision department can take good control of such crowded situations in the future.
News:	1. 昨日，有多家超市发通知称，为吸取上海踩踏事故教训，取消门店促销活动 Yesterday, many supermarkets make announcements to cancel their store promotions so as to take lessons from the Shanghai Stampede accident. 2. 在人员密集场所发生踩踏事件时掌握一定的自救知识很重要。 It is important to learn some knowledge about self rescue when a stampede breaks out at a crowded place. 3. 如果身陷一个人潮汹涌、进退不得的人群之中，为了避免发生踩踏事故，最好的自救方法就是联合你前后左右的人一起采用人体麦克法，有节奏的呼喊：“后退”。 If you stand in a thick crowd and can not move, the best way of rescuing yourself is shouting fallback rhythmically together with people surrounding you. 4. 为避免再次发生事故，上海警方在现场摆上护栏、筑起人墙，以控制人流行进的方向和速度。 In order to avoid the accident from happening again, Shanghai police has setup the guardrail and lined up a wall to control the moving direction and speed of the crowd. 5. 由于天气晴好，事发现场的警戒线已经撤除，台阶又恢复了往日繁忙的景像，外滩依旧人山人海 Due to the fine weather, the guard line on the scene has been removed, the stairs restore its busy sight, and the Bund is crowded with people as usual.

Fig. 4. Contrastive opinions from news & tweets towards common aspect 3.

outbreak of the stampede and hence saved many lives. The common aspect 2 is about the announcement of the victims namelist in this stampede, while the common aspect 3 is to advocate public security sense and discuss precautions during a stampede. These results well demonstrate the ability of CAMEL in learning high-quality common aspects, especially for the tweet side; which is typically unclear and contains fewer aspect words. Note that there are some other common aspects concerning the topics such as *rescue*, *penalty*, *etc*., induced by CAMEL. We omit them here to avoid redundancy.

We then take a close look at the contrastive opinions learned by CAMEL towards the common aspects in Fig. 3, where “opinion 0” and “opinion 1” mean the opinion words extracted from tweets and news, respectively. In general, the opinions from the tweet side seem obviously more emotional than those from news. For example, from the tweet opinions towards the common aspect 1, we find the public is greatly touched by the fallback shouters and appreciates their sensible behavior, which however is much less obvious for the opinion words from news. Similar situations happen for the opinions towards the common aspect 2, where on the tweet side most people feel sad with those dead young, while some criticize them for their ignorance. But on the news side, the opinion words seem much more formal and monotonic with only the mourning sentiment. For the common aspect 3, the first glance of the two sets of opinion words appears indistinguishable. However, if we further select, randomly, five sentences assigned to this common aspect from the two sides, respectively (see Fig. 4), we find the subtle differences between the two opinions. That is, while all sentences in the tweets and news are talking about the precaution of the stampede, tweets are mainly talking about *self precaution* and wish the government to exercise better precaution, and the news are mainly talking about the efforts the government made to prevent the recurrence of a stampede. Obviously, the tweets express much stronger emotions than the news.

Finally, Figs. 5 and 6 illustrate the specific aspects and opinions from tweets and news, respectively. As can be seen from Fig. 5, the public talks a lot about the improvements of national quality and the public security awareness. While in Fig. 6, we find the news reports the lessons the government should draw from the stampede, which are different from some tweets that care more about the possible punishment of the persons in charge.

7 Related Work

7.1 Aspect-Based Opinion Mining

Two subtasks are usually involved in this problem, namely, *aspect or feature identification* and *opinion extraction*. Most of the early works on aspect identification are feature-based approaches [10], [21], e.g., applying frequent itemset mining to identify product aspects [17], which normally exert some constraints on high-frequency noun phrases to find aspects. As a result, they are usually subject to the risk of producing too many non-aspects examples and missing low-frequency aspects [8]. Several early works have applied supervised learning to identify both aspects and opinions [11], [12], [33], which, however, needs hand-labeled training sentences and thus is very costly.

In recent years, with the popularity of topic models, more unsupervised methods are proposed for aspect-based opinion mining. For instance, Titov and McDonald propose a multi-grain topic model to learn both global and local topics, in which local topics correspond to rateable aspects [29]. Another approach to discover aspects is to fit a topic model to sentences instead of documents. For instance, Brody and Elhadad run the latent dirichlet allocation(LDA) [2] model over sentences instead of documents to extract aspects [3]. Zhao et al. [35] and Jo et al. [13] assume that all words in a single sentence are generated from one topic.

Some researchers take approaches that model topic and sentiment in a unified way. For instance, Lin and He

Aspect	素质(competence) 提高(improvement) 国民(citizen) 有待(needs) 国人(countrymen) 安全(safety) 素养(attainment) 民众(the_public) 秩序(order) 国家(nation) 意识(awareness)
Opinion	提高(improve) 调查(investigate) 客观(objective) 需要(need) 文明(civilized) 宽慰(comfort) 安全(safe) 教育(educate) 偶然(accidental) 无序(unordered)

Fig. 5. Sample specific aspect and opinion from tweets.

Aspect	事件(event) 教训(lesson) 上海(Shanghai) 跨年(new_year) 事故(accident) 原因(reason) 政府(government) 报告(report) 生者(survivor) 台阶(stairs)
Opinion	调查(investigate) 客观(objective) 处理(dispose) 公布(announce) 严查(strictly_investigate) 尊重(respect) 真实(true) 宽慰(comfort) 关注(care) 认真(serious)

Fig. 6. Sample specific aspect and opinion from news.

propose a joint topic-sentiment (JTM) model to detect sentiment and topic simultaneously from texts [15]. The aspect and sentiment unification model (ASUM) proposed by Jo et al. [13] is similar to JTM; the major difference lays in that ASUM assumes each single sentence only covers one topic. The above two models do not explicitly separate topic words and sentiment words. Mei el al. [19] propose a topic-sentiment mixture model, which represents positive and negative sentiments as language models separating from topics, but both models only capture general opinion words. Brody et al. [3] take a two-step approach by first detecting aspects and then identifying aspect-specific opinion words. Zhao et al. [35] propose a topic model integrating with a maximum entropy model (MaxEnt-LDA) to jointly capture both aspects and aspect-specific opinion words within a topic model. Detailed discussions about aspect-specific opinion models based on LDA can be found in [23].

7.2 Cross-Collection Text Mining (CCTM)

7.2.1 Parametric CCTM

Zhai et al. [34] introduce a task called "comparative text mining" and propose a cross-collection mixture (ccMix) model based on probabilistic latent semantic index (pLSA) [9]. The goal of the task is to discover the common themes across all collections and the ones unique to each collection. Paul et al. [26] extend ccMix to the ccLDA model based on LDA for cross-culture topic analysis. Gao et al. [6] propose a cross-collection topic aspect model (ccTAM) to perform event summarization across news and social media steams. They assume aspects contained only in tweets can serve as a supplement to those in the news. Fang et al. [5] propose a cross-perspective topic model (CPT) to perform contrastive opinion modeling. They view opinions with the same topic yet from different news sources as different perspectives, and learn topics (not *aspects*) across collections by performing LDA over aggregated collections. CPT has no guarantee to find shared topics, especially when collections are less comparative, such as news versus tweets in our case.

7.2.2 Nonparametric CCTM

A major limitation of parametric models is that they require to specify a fixed number of topics as a prior, which is usually very hard in practice. Bayesian nonparametric models, especially those based on Dirichlet processes (DP), are popularly adopted to address the above issue. Hierarchical Dirichlet process [28] is one of the most popular nonparametric topic model, and a two-level HDP can seem to be like an "infinite LDA". HDP of three levels is used to model multiple corpora, which often results in better performance than HDP with two levels. However, a three-level HDP cannot distinguish common and specific topics over collections. Muller et al. [24] suggest using linear combinations of realizations of independent DPs to achieve dependence among random measures. This approach indeed inspires our CAMEL-DP, where a group-specific DP is a linear combination of a global DP and a local one, modeling the same concept of common/specific aspects as CAMEL. Lin et al. [16] introduce a more general framework, namely coupled nonparametric mixture model, to couple latent Dirichlet Processes, and our CAMEL-DP is also built on Lin's framework. Other related works include the hybrid nested hierarchical Dirichlet process proposed by Ma et al. [18], which models common and specific topics across document clusters (not observed document collections).

8 Conclusions

In this paper, we proposed CAMEL, a novel topic model for complementary aspect-based opinion mining across asymmetric collections. By modeling both common and specific aspects while keeping contrastive opinions, CAMEL is capable of integrating complementary information from different collections in both aspect and opinion levels. An auto-labeling scheme called AME with word embedding-based similarity enhancements was also introduced to further allow CAMEL to suit real-life applications. Moreover, a nonparametric alternative to CAMEL called CAMEL-DP was also proposed based on coupled Dirichlet Processes to avoid the dilemma of setting a proper topic number. Extensive experiments and a real-world case study on a public event demonstrated the effectiveness of CAMEL and CAMEL-DP in leveraging collection complementarity for high-quality aspect and opinion mining. In the future work, we would like to explore whether the AME scheme can adapt to all types of opinionated texts.

Acknowledgments

Dr. Junjie Wu was supported by the National Natural Science Foundation of China (NSFC) (71531001, 71725002, U1636210, 71471009, 71490723, 71322104, 71171007), and Fundamental Research Funds for Central Universities. Dr. Deqing Wang was supported by NSFC (71501003) and the China Postdoctoral Science Foundation funded project (2014M550591). A preliminary version of this manuscript has been published as a full conference paper in ICDM'15 [36].

References

[1] Y. Bao, N. Collier, and A. Datta, "A partially supervised cross-collection topic model for cross-domain text classification," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2013, pp. 239–248.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[3] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Proc. Human Language Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 804–812.

[4] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Proc. Human Language Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 804–812.

[5] Y. Fang, L. Si, N. Somasundaram, and Z. Yu, "Mining contrastive opinions on political texts using cross-perspective topic model," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 63–72.

[6] W. Gao, P. Li, and K. Darwish, "Joint topic modeling for event summarization across news and social media streams," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1173–1182.

[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Academy Sci. United States America*, vol. 101, pp. 5228–5235, 2004.

[8] H. Guo, H. Zhu, Z. Guo, X. Zhang, and Z. Su, "Product feature categorization with multilevel latent semantic association," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 1087–1096.

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168–177.

[11] W. Jin and H. H. Ho, "A novel lexicalized HMM-based learning framework for web opinion mining," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 465–472.
[12] W. Jin, H. H. Ho, and R. K. Srihari, "OpinionMiner: A novel machine learning system for web opinion mining and extraction," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 1195–1204.
[13] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 815–824.
[14] K. W. Lim and W. Buntine, "Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 1319–1328.
[15] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 375–384.
[16] D. Lin and J. W. Fisher, "Coupling nonparametric mixtures via latent Dirichlet processes," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 55–63.
[17] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *Proc. 14th Int. Conf. World Wide Web*, 2005, pp. 342–351.
[18] T. Ma, I. Sato, and H. Nakagawa, "The hybrid nested/hierarchical Dirichlet process and its application to topic modeling with word differentiation," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2835–2841.
[19] Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 171–180.
[20] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 262–272.
[21] S. Moghaddam and M. Ester, "Opinion digger: An unsupervised opinion miner from unstructured product reviews," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1825–1828.
[22] S. Moghaddam and M. Ester, "Aspect-based opinion mining from product reviews," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 1184–1184.
[23] S. Moghaddam and M. Ester, "On the design of LDA models for aspect-based opinion mining," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 803–812.
[24] P. Müller, F. A. Quintana, and G. Rosner, "A method for combining inference across related nonparametric Bayesian models," *J. Roy. Statistical Soc. Series B*, vol. 66, no. 3, pp. 735–749, 2004.
[25] S. Park, S. J. Lee, and J. Song, "Aspect-level news browsing: Understanding news events from multiple viewpoints," in *Proc. 15th Int. Conf. Intell. User Interfaces*, 2010, pp. 41–50.
[26] M. Paul and R. Girju, "Cross-cultural analysis of blogs and forums with mixed-collection topic models," in *Proc. Conf. Empirical Methods Natural Language Process.: Volume 3*, 2009, pp. 1408–1417.
[27] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn.*, 2012, pp. 952–961.
[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statistical Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.
[29] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 111–120.
[30] J. Wang, et al., "Mining multi-aspect reflection of news events in Twitter: Discovery, linking and presentation," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 429–438.
[31] R. Wang, W. Huang, W. Chen, T. Wang, and K. Lei, "ASEM: Mining aspects and sentiment of events from microblog," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1923–1926.
[32] Y. Wu and M. Ester, "FLAME: A probabilistic model combining aspect based opinion mining and collaborative filtering," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 199–208.
[33] Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining," in *Proc. Conf. Empirical Methods Natural Language Process.: Volume 3*, 2009, pp. 1533–1541.
[34] C. X. Zhai, A. Velivelli, and B. Yu, "A cross-collection mixture model for comparative text mining," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 743–748.
[35] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2010, pp. 56–65.
[36] Y. Zuo et al., "Complementary aspect-based opinion mining across asymmetric collections," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 669–678.

Yuan Zuo received the PhD degree from Beihang University, Beijing, China, in 2017. He is currently a postdoctor in the Information Systems Department, Beihang University. His research interests include topic modeling and social computing.

Junjie Wu received the PhD degree in management science and engineering from Tsinghua University. He is currently a full professor in the Information Systems Department, Beihang University, and the director of the Beihang Social Computing Center. His general area of research is data mining and complex networks. He is the recipient of the NSFC Distinguished Young Scholars award and MOE Changjiang Young Scholars award in China.

Hui Zhang received the MS and PhD degrees in computer science from Beihang University, China, in 1994 and 2009, respectively. He is currently a professor in the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University. Now, he is working for the Chinese Science and Technology Resources Portal (http://www.escience.gov.cn) as a chief architect. His main research interests include cloud computing, web information retrieval, and data mining.

Deqing Wang received the doctoral degree in computer science from Beihang University, in 2012. He is an assistant professor in the School of Computer, Beihang University. Before that, he was a post-doctoral fellow in the School of Economics and Management, Beihang University, China. His research focuses on text categorization, data mining for software engineering, and machine learning.

Ke Xu received the BE, ME, and PhD degrees from Beihang University, in 1993, 1996, and 2000, respectively. He is a professor with Beihang University, China. His research interests include algorithm and complexity, data mining, and complex networks.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.