

文章编号: 129

基于联合模型的商品口碑数据情感挖掘*

李大宇¹, 李旻¹, 王素格^{1,2}

(1.山西大学 计算机与信息技术学院, 山西 太原 030006;

2.山西大学 计算智能与中文信息处理教育部重点实验室, 山西 太原 030006)

摘要: 为了同时挖掘商品口碑数据中所谈论的对象、对象的某个方面以及评论者对这个方面的观点, 用于指导消费者的消费和生产厂家对商品的改进, 本文面向口碑数据提出一个无监督对象方面情感联合模型。该模型假设方面分布依赖于对象分布, 情感分布依赖于方面分布和对象分布, 词是采样的最小单位。在汽车口碑数据上进行了多组实验, 实验表明, 无监督对象方面情感联合模型不仅可以判别文本方面和文本情感的类别, 而且还可以获取文本对象信息。

关键词: 口碑数据; 无监督; 对象方面情感联合模型

中图分类号: TP391

文献标识码: A

Sentiment Mining of Product Reputation Data Based on Joint Model

Li Dayu^{1,2}, Li Yang¹, Wang Suge^{1,2}

(1.School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China; 2.Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: Text mining aims to discover hidden knowledge automatically from text. In this paper, we propose a novel probabilistic graphical model, called joint object-aspect-sentiment model, which extracts objects, aspects and sentiments simultaneously from product reputation data. Unlike other aspect-based topic model approaches which detecting only aspects and sentiments, the proposed joint object-aspect-sentiment model not only captures aspects and sentiments information but also objects from reviews, besides it is fully unsupervised. The underlying assumption of joint object-aspect-sentiment model is that aspect distribution depends on object distribution, sentiment distribution depends on aspect distribution. It further assumes that words are the smallest units of sampling. Our experiments on the car reputation data show that joint object-aspect-sentiment model can predict the aspect and sentiment categories of product review texts, and additionally extracts objects information from reviews simultaneously.

Key words: Product reputation data; Topic model; Joint object-aspect-sentiment model

1 引言

随着电子商务的蓬勃发展, 越来越多的人喜欢在微博和论坛上对商品进行评论, 致使网络上存在着大量商品口碑数据。所谓的口碑数据是指消费者在购买商品后, 对所购商品某个具体特性的观点评述。这些口碑数据中包含很多对消费者和生产厂家有用的信息, 例如“奥迪 A6 的空间很宽敞”, 评论中的评论对象是“奥迪 A6”, 谈论的是它的“空间”方面, 其观点是“宽敞”。对于同一种商品, 存在多种品牌, 例如汽车, 常见的品牌有“奥迪”、“大众”、“现代”等, 我们可以通过“空间”、“动力”、“油耗”等方面来评价其中一款汽车的好坏。因此, 同时从数据中挖掘出消费者关注的商品品牌、商品的方面以及对方面的观点具有重要的应用价值。一方面, 消费者对这款商品的某个方面在做出是否购买这款商品的决定时

* 收稿日期:

定稿日期:

基金项目: 国家自然科学基金资助项目(61573231, 61272095, 61432011, U1435212, 61175067), “863” 高技术研究计划基金项目 (2015AA015407), 山西省科技基础条件平台计划项目 (2015091001-0102), 山西省回国留学人员科研项目(2013-014)

其他消费者的观点将起到关键作用；另一方面，商品生产厂家也需要同其他品牌的商品进行对比以便改进某些方面。在生活节奏和商品更新换代速度日益加快的今天，通过人工逐条阅读商品评论耗时耗力。因此，如何从海量评论数据中迅速有效的挖掘出评论的对象、方面和情感尤为重要。

近年来，有大量基于主题模型方法的商品评论文本挖掘的报道^[2-12]，但大都是文本-方面(主题)-情感-词四层模型，并没有加入对象信息，这些模型无法获取评论文本谈及的对象。从模型的依赖性上 有的假设情感分布依赖于方面分布^[2-6]，有的则假设方面分布依赖于情感分布^[7,8]。在应用中，它们更加关注文本情感分类而忽略模型对文本方面分类影响。直观上，情感极性依赖于对象和方面，相同情感词在谈论的对象或者方面不同时，表达的情感也不同，例如，在谈论汽车时，“大”在“油耗大”中表达的是消极情感，在“空间大”中表达的是积极情感，所以同时挖掘对象、方面和情感对文本观点挖掘显得尤为重要。

针对这些问题，本文在 LDA 模型 (Latent Dirichlet Allocation) 的基础上，提出了无监督对象方面情感联合模型(JOAS)。与以往的模型不同，该模型采用文本-对象-方面-情感-词五层生成式结构，并假设方面分布依赖于对象分布，情感分布依赖于方面分布和对对象分布。模型采用完全无监督学习方法，并且可以同时从文本中获取对象、方面和情感信息。对于评论文本中每一个词，首先从文本-对象分布中选择一个谈论的对象标签，然后从对象-方面分布中选择对象的一个方面，再从方面-情感分布中选择出对这个方面的情感。通过构建了 JOAS，可以用于指导消费者购买哪种品牌或者型号的商品。

本文组织为第二部分主要列出主题模型在对象方面情感挖掘的相关研究；第三部分介绍了本文提出的 JOAS 模型；第四部分介绍了实验过程的建立；第五部分则讲述了模型用于对象分类、方面分类和情感分类及其联合和方面情感词获取的实验结果及其详细分析；第六部分是对本文提出方法的一个简要总结和对未来工作的展望。

2 相关工作

2003 年，Blei 等人^[1]借鉴概率图模型的理论和方法，提出了 LDA 模型，该模型是一个完整的概率生成模型，参数空间的规模与文本数量无关，适合处理大规模的语料库。由于 LDA 模型具有良好的数学基础以及灵活的拓展性，越来越多的研究将其拓展并应用到方面情感联合分析中。

假设方面分布依赖于情感分布，Lin 等^[2-4]提出了 JST 模型(joint sentiment/topic model)及其扩展，Jo 等^[5]提出了 ASUM 模型(aspect and sentiment unification model)和孙艳等^[6]提出了 UTSU 模型(unsupervised topic and sentiment unification model)。JST 模型采样主题标签和情感标签时，对每个词采样主题标签和情感标签；ASUM 模型采样主题标签和情感标签时，对每个句子进行采样；UTSU 模型则是对每个句子采样情感标签，对每个词采样主题标签。

假设情感分布依赖于方面分布，Li 等^[7,8]提出了 Sentiment-LDA 模型，该模型采样主题标签和情感标签时，对每个词采样主题标签和情感标签。

以上这些模型，一方面他们都是四层结构，没有加入对象信息，另一方面都侧重于文本情感分类而忽略文本方面分类。

此外还有大量有监督的模型用于方面情感分析^[9-18]，Mei 等人^[9]提出了一个主题情感混合模型(Topic-Sentiment Mixture, TSM)，但他们认为主题分布和情感分布是独立的，并没有反映主题和情感间的相互影响；Titov 等人^[10-11]提出了一种多粒度 LDA 模型(Multi-Grain LDA, MG-LDA)，用以挖掘产品领域情感文本中评价对象，并将相似的评价对象进行聚类；Zhao 等人^[12]提出的 ME-LDA 模型(MaxEnt-LDA)，该模型结合了最大熵组件和主题模型。这些有监督的方法不足是大量的样本不容易获取和领域移植性差。

3 JOAS 模型

在一条口碑数据中，假设只涉及一款汽车，关于这款汽车的一个方面消费者对其发表自己的观点。例如：

“这个相信关注过福克斯的人都应该知道，空间是小福的弱点，后排确实有些局促。”

“新福克斯将近 1.4 吨的自重配上 1.6L 的发动机稍显够呛，在起步阶段可感觉发动机动力不足，夏天开空调更是如此。”

为了获取口碑数据中的汽车对象、方面和情感，原有的 LDA 模型是文本-主题-词三层结构模型，仅有一个主题层是隐层，在此基础上，本文通过添加对象层和情感层，构建了对对象方面情感联合模型(JOAS)。假定情感标签依赖于方面标签和对象标签，方面标签依赖于对象标签。

3.1 JOAS 模型的生成过程

假定语料库 C 为 D 篇文本的集合，记为 $C = \{d_1, d_2, \dots, d_D\}$ ；文本 d 存在 N_d 个词，记为 $d = (w_1, w_2, \dots, w_{N_d})$ ；每个词 w_i 均是词表中的一项（词表由 V 个不同的词构成： $\{1, 2, \dots, V\}$ ）；设 T 为对象的个数， J 为方面的个数， K 为情感极性的个数。则 JOAS 模型对应的文本生成过程如下（ $\alpha, \beta, \gamma, \eta$ 为模型超参数， $Dir()$ 为狄利克雷分布， $Multi()$ 为多项分布）：

- 对每一篇文本 d ，选取一个文本-对象分布 $\lambda_d \sim Dir(\eta)$ ；
- 对于文本 d 下的对象 o 选取对象-方面分布 $\theta_{d,o} \sim Dir(\gamma)$ ；
- 对于文本 d 对象 o 下的方面 a 选取对象-方面-情感分布 $\pi_{d,o,a} \sim Dir(\alpha)$ ；
- 对于文本 d 中的每一个词 w_i ：
 - 选取对象标签 $o_i \sim Multi(\lambda_d)$ ；
 - 选取方面标签 $a_i \sim Multi(\theta_{d,o_i})$ ；
 - 选取情感标签 $s_i \sim Multi(\pi_{d,o_i,a_i})$ ；
 - 选取词 $w_i \sim Multi(\varphi_{o_i,a_i,s_i})$ ，其中 $\varphi_{o_i,a_i,s_i} \sim Dir(\beta)$ 为词 w_i 在对象方面情感下的概率分布。

对应上述生成过程，构建 JOAS 模型图如图 1 所示。

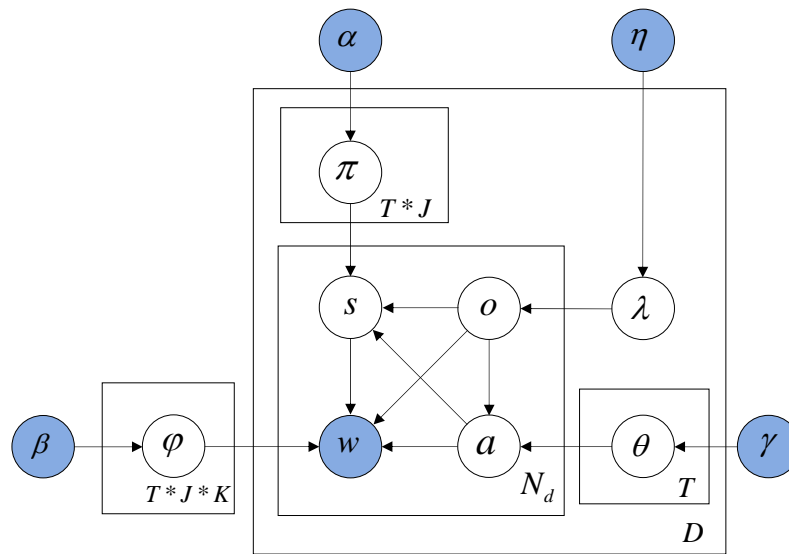


图 1 JOAS 模型图

3.2 JOAS 模型参数估计

JOAS 模型有四组多项分布参数，分别是文本-对象分布 Λ ，文本-对象-方面分布 Θ ，文本-对象-方面-情感分布 Π 和对象-方面-情感-词分布 Φ 。为了获取这四组参数，我们进行了以下推导。

3.2.1 联合概率

语料库中所有词及其情感标签、方面标签、主题标签的联合概率可以分解为如下形式：

$$\begin{aligned} P(w, s, a, o) &= P(w | s, a, o) P(s, a, o) \\ &= P(w | s, a, o) P(s | a, o) P(a, o) \\ &= P(w | s, a, o) P(s | a, o) P(a | o) P(o) \end{aligned} \quad (1)$$

其中， w 为语料库中所有词的记号序列， o 为语料库中所有词的对象标签序列， a 为语料库中所有词的方面标签序列， s 为语料库中所有词的情感标签序列，假定语料库中词与词之间是相互独立的。

公式(1)中的 $P(w | s, a, o)$ 、 $P(s | a, o)$ 、 $P(a | o)$ 和 $P(o)$ 通过分别对 Φ 、 Π 、 Θ 和 Λ 积分可以得到：

$$P(w | s, a, o) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^{T*J*K} \cdot \prod_{t=1}^T \prod_{j=1}^J \prod_{k=1}^K \frac{\prod_{v=1}^V \Gamma(N_{t,j,k,v} + \beta)}{\Gamma(N_{t,j,k} + V\beta)} \quad (2)$$

$$P(s | a, o) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \right)^{D*T*J} \cdot \prod_{d=1}^D \prod_{t=1}^T \prod_{j=1}^J \frac{\prod_{k=1}^K \Gamma(N_{d,t,j,k} + \alpha)}{\Gamma(N_{d,t,j} + K\alpha)} \quad (3)$$

$$P(a | o) = \left(\frac{\Gamma(J\gamma)}{\Gamma(\gamma)^J} \right)^{D*T} \cdot \prod_{d=1}^D \prod_{t=1}^T \frac{\prod_{j=1}^J \Gamma(N_{d,t,j} + \gamma)}{\Gamma(N_{d,t} + J\gamma)} \quad (4)$$

$$P(o) = \left(\frac{\Gamma(T\eta)}{\Gamma(\eta)^T} \right)^D \cdot \prod_{d=1}^D \frac{\prod_{t=1}^T \Gamma(N_{d,t} + \eta)}{\Gamma(N_d + T\eta)} \quad (5)$$

其中， V 是词表的大小， T 是对象标签的总数， J 是方面标签的总数， K 是情感标签的总数， D 是语料中文档总数。 $N_{t,j,k,v}$ 为语料库中对象标签为 t 、方面标签为 j 、情感标签为 k 的词 v 的个数； $N_{d,t,j,k}$ 为文本 d 中对象标签为 t 、方面标签为 j 、情感标签为 k 的词的个数； $N_{d,t,j}$ 为文本 d 中对象标签为 t 、方面标签为 j 的词的个数； $N_{d,t}$ 为文本 d 中对象标签为 t 的词的个数； N_d 为文本 d 中的词汇总数。

3.2.2 Gibbs sampling^[19]

为了得到模型参数，我们用 Gibbs 采样算法对模型的变量 o_i 、 a_i 和 s_i 进行采样，利用 3.2.1 节所得联合概率，给定除当前词 w_i 的对象、方面和情感标签外的标签，语料库中词 w_i 的对象标签 o_i 为 t ，方面标签 a_i 为 j ，情感标签 s_i 为 k 的概率为：

$$\begin{aligned} & p(s_i = k, a_i = j, o_i = t | s_{-i}, a_{-i}, o_{-i}, w, \alpha, \beta, \gamma, \eta) \\ & \propto \frac{\left\{ N_{t,j,k,v} \right\}_{-i} + \beta}{\left\{ N_{t,j,k} \right\}_{-i} + V\beta} \cdot \frac{\left\{ N_{d,t,j,k} \right\}_{-i} + \alpha}{\left\{ N_{d,t,j} \right\}_{-i} + K\alpha} \cdot \frac{\left\{ N_{d,t,j} \right\}_{-i} + \gamma}{\left\{ N_{d,t} \right\}_{-i} + J\gamma} \cdot \frac{\left\{ N_{d,t} \right\}_{-i} + \eta}{\left\{ N_d \right\}_{-i} + T\eta} \end{aligned} \quad (6)$$

其中， o_{-i} 为除第 i 个词外的其他词汇的对象标签序列； a_{-i} 为除第 i 个词外的其他词汇的方面标签序列； s_{-i} 为除第 i 个词外的其他词汇的情感标签序列； $\left\{ N_{t,j,k,v} \right\}_{-i}$ 为语料库中除第 i

个词外，对象标签为 t 、方面标签为 j 、情感标签为 k 的词 v 的个数； $\{N_{t,j,k}\}_{-i}$ 为语料库中除第 i 个词外，对象标签为 t 、方面标签为 j 、情感标签为 k 的词的个数； $\{N_{d,t,j,k}\}_{-i}$ 为文本 d 中除第 i 个词外，对象标签为 t 、方面标签为 j 、情感标签为 k 的词的个数； $\{N_{d,t,j}\}_{-i}$ 为文本 d 中除第 i 个词外，对象标签为 t 、方面标签为 j 的词的个数； $\{N_{d,t}\}_{-i}$ 为文本 d 中除第 i 个词外，对象标签为 t 的词的个数； $\{N_d\}_{-i}$ 为文本 d 中除第 i 个词外的词汇总数。

3.2.3 参数估计

当 Gibbs sampling 收敛后，我们需要根据最后文档集中所有词的对象标签、方面标签和情感标签分配情况来计算参数矩阵 Λ 、 Θ 、 Π 和 Φ 。四组多项分布参数：文本-对象分布 Λ 、文本-对象-方面分布 Θ 、文本-对象-方面-情感分布 Π 和对象-方面-情感-词分布 Φ 的估计公式如下：

$$\lambda_{d,t} = \frac{N_{d,t} + \eta}{N_d + T\eta} \quad (7)$$

$$\theta_{d,t,j} = \frac{N_{d,t,j} + \gamma}{N_{d,t} + J\gamma} \quad (8)$$

$$\pi_{d,t,j,k} = \frac{N_{d,t,j,k} + \alpha}{N_{d,t,j} + K\alpha} \quad (9)$$

$$\varphi_{t,j,k,v} = \frac{N_{t,j,k,v} + \beta}{N_{t,j,k} + V\beta} \quad (10)$$

3.3 基于 JOAS 模型的标签识别

对于模型 JOAS，利用公式(7)中的 $\lambda_{d,t}$ 可以得到对象 t 在文本 d 中的概率估计，通过公式(8)中的 $\theta_{d,t,j}$ 可以得到方面 j 在文本 d 对象 t 中概率估计，通过公式(9)中的 $\pi_{d,t,j,k}$ 可以得到情感 k 在文本 d 对象 t 方面 j 中概率估计。因此，利用这些参数估计，JOAS 模型可采用公式(11)-(13)对文本对象类别、方面类别以及文本的情感类别进行判断。

$$o_d = \underset{t}{\operatorname{argmax}} \left\{ \lambda_{d,t} \mid t \in [1, 2, \dots, T] \right\} \quad (11)$$

$$a_d = \underset{j}{\operatorname{argmax}} \left\{ \sum_{t=1}^T \lambda_{d,t} \cdot \theta_{d,t,j} \mid j \in [1, 2, \dots, J] \right\} \quad (12)$$

$$s_d = \underset{k}{\operatorname{argmax}} \left\{ \sum_{t=1}^T \sum_{j=1}^J \lambda_{d,t} \cdot \theta_{d,t,j} \cdot \pi_{d,t,j,k} \mid k \in [1, 2, \dots, K] \right\} \quad (13)$$

其中 o_d 为文本 d 的对象标签、 a_d 为文本 d 的方面标签和 s_d 为文本 d 的情感标签。

利用公式(10)中的分布 $\varphi_{t,j,k,v}$ 可以得到在对象 t 方面 j 情感 k 下词 v 的概率估计，见公式 (14) 和 (15)。

$$p_{j,v} = \sum_{t=1}^T \sum_{k=1}^K \varphi_{t,j,k,v} \quad (14)$$

$$p_{j,k,v} = \sum_{t=1}^T \varphi_{t,j,k,v} \quad (15)$$

其中， $p_{j,v}$ 为方面-词的概率估计和 $p_{j,k,v}$ 为方面-情感-词的概率估计。

4 实验过程的建立

为了验证 JOAS 模型的有效性，建立如下的实验过程：

(1) 采用 ICTCLAS 开源工具包对实验数据进行分词，并将分词后的文本剔除停用词；
(2) 利用文献[20]所构建的汽车评价本体知识库初始化对象标签、方面标签和情感标签。这些先验知识仅在模型初始化时用到。

(3) 根据 3.2.2 节的 Gibbs sampling 算法进行采样，迭代一定次数后模型各个参数趋于稳定；

(4) 根据公式(11)、(12)和(13)可以得到文本的对象、方面和情感标签，分别与事先人工标好的标签进行比对，计算文本分类正确率。

(5) 根据公式(14)和(15)，从汽车评论数据中抽取概率排名在前面的<方面,词>和<方面,情感,词>；

实验评价指标采用正确率 *Accuracy*，计算公式如下：

$$Accuracy = \frac{N_{same}}{N_{total}} \quad (16)$$

其中， N_{same} 为真实标签和模型计算的标签相同的文本数， N_{total} 为数据集总文本数。

5 实验结果

5.1 实验数据及预处理

本文实验数据为汽车之家网(<http://www.autohome.com.cn/>)上下载的汽车口碑数据，由于紧凑型车是当下最流行的家庭用车，其销售量在家庭用车中也最高，因此，实验选取紧凑型车里评论热度最高的前四种汽车品牌：福克斯、科鲁兹、凯越和逸动，这些品牌裸车价格在 7 万到 15 万之间，价格适中，购买量大。并且选取用于评价汽车的五个方面：空间、动力、油耗、舒适性和内饰。手工标注了每篇评论的极性用于测试，选取正负各 3741 篇，整理共得到 7482 篇评论，见表 1 所示。

表 1 实验数据集

车型	空间		动力		油耗		舒适性		内饰		累计 正负
	正	负	正	负	正	负	正	负	正	负	
福克斯	261	261	168	168	127	127	249	249	181	181	1972
科鲁兹	231	231	269	269	143	143	224	224	213	213	2160
凯越	165	165	193	193	153	153	169	169	118	118	1596
逸动	118	118	218	218	109	109	236	236	196	196	1754
累计	1550		1696		1064		1756		1416		7482

5.2 实验设置

JOAS 模型旨在同时从文本中获取对象、方面和情感信息，为了评价 JOAS 模型，在汽车评论数据集上进行了三组实验：

- (1) 模型迭代次数的确定。
- (2) 文本对象分类、文本方面分类和文本情感分类比较实验。
- (3) 方面-词获取，方面-情感-词获取。

5.3 迭代次数的确定

本文采用 Gibbs sampling 算法进行采样，为了确定模型采样的迭代次数，我们将 α 、 β 、 γ 和 η 均固定为 0.01，观察迭代次数对文本情感分类的影响，结果如图 2 所示。

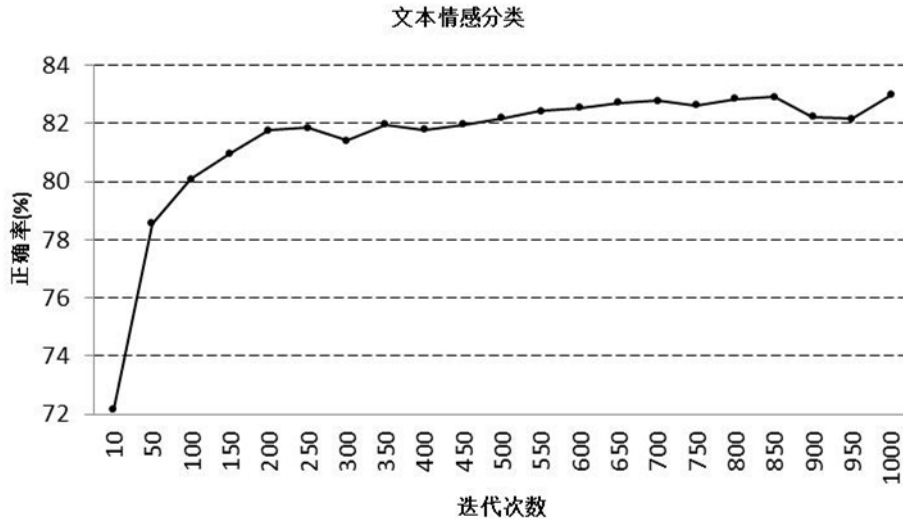


图2 不同迭代次数下的文本情感分类正确率

通过图2可以看出,迭代次数为10-200时,分类正确率随迭代次数的增加而快速提升,当迭代次数大于200后,模型分类结果趋于稳定,这与Gibbs sampling算法的特性是一致的。综合考虑程序运行时间和结果稳定性,选取模型迭代次数800,后续的实验也采用此值。

5.4 不同参数下模型间实验的比较结果

为了验证本文JOAS模型的有效性,将其与文献[2]提出的JST模型比较。对于JOAS模型,包含4个超参数 α 、 β 、 γ 和 η ,而对比模型JST仅包含3个超参数 α 、 β 和 γ 。由于超参数作为模型的先验知识对结果有一定的影响,而参数的组合情况比较多,因此根据文献[2],将每个参数分别设置为0.01和0.1。为了考察不同参数对模型分类的影响,根据公式(11)-(13),对文本的对象分类、方面分类和情感分类及其联合分别进行实验,得出文本的对象类别、方面类别和情感标签类别及其联合的正确率。

由于JOAS模型是无监督学习方法,为了比较与有监督学习方法的差异,本文选择SVM分类器作为有监督学习方法的代表,其采用的文本特征为bag-of-words特征,并采用十折交叉验证。SVM分类器将对象、方面、情感分类视为独立无关的分类任务。

JOAS模型、文献[2]提出的JST模型以及SVM分类器方法的比较实验结果如表2所示。

表2 在不同参数下的文本对象分类、方面分类和情感分类及其联合分类正确率(%)

	参数				对象 分类	方面 分类	情感 分类	<对象, 情感>	<方面, 情感>	<对象,方 面,情感>
	α	β	γ	η						
JOAS	0.01	0.01	0.01	0.01	69.93	92.80	81.52	57.15	76.34	53.81
	0.1	0.01	0.01	0.01	71.63	93.93	80.82	58.23	76.25	55.21
	0.01	0.1	0.01	0.01	77.76	94.63	75.46	58.50	71.59	55.91
	0.01	0.01	0.1	0.01	68.50	91.17	78.86	53.93	72.47	49.61
	0.01	0.01	0.01	0.1	52.21	91.51	82.01	43.00	75.55	39.25
JST	0.01	0.01	0.01	-	-	92.54	80.49	-	74.79	-
	0.1	0.01	0.01	-	-	89.80	82.41	-	74.06	-
	0.01	0.1	0.01	-	-	94.79	79.43	-	75.31	-
	0.01	0.01	0.1	-	-	93.79	79.46	-	74.61	-
SVM	-	-	-	-	41.15	83.68	76.33	-	-	-

由表2可以看出:

- (1) 不同的参数设置对应的模型结果不相同

对于 JOAS 模型来说, 当四个参数组合为(0.01, 0.1, 0.01, 0.01)时, 文本对象分类正确率和文本方面分类正确率均达到最高, 分别为 77.76%和 94.63%, 而对应的情感分类正确率为 75.46%; 当参数组合为(0.01, 0.01, 0.01, 0.1)时, 文本情感分类正确率达到最高的 82.01%, 此时, 文本对象分类的正确率为 52.21%, 文本的方面分类正确率为 91.51%。对于 JST, 参数为(0.01, 0.1, 0.01), 文本方面分类达到最高 94.79%, 而文本情感分类为 79.43%; 参数为(0.1, 0.01, 0.01)时, 文本情感分类达到最高的 82.41%, 文本方面分类为 89.80%。

(2) JOAS 模型与 JST 模型比较结果

①在方面分类、文本情感分类方面, 两者模型的结果相当。主要原因是本文实验选取的语料都是汽车领域的, 同一领域内不同汽车品牌之间所采用的方面词或情感词之间差异不大, 导致在方面分类和情感分类上模型间分类能力相当。

②两个模型在文本方面分类上准确率较高, 均达到 90%以上, 主要由于本文实验采用的是汽车口碑数据, 对汽车某个方面的评论描述较为细致。

③在<方面,情感>联合实验中, JOAS 模型略优于 JST 模型。另外, JOAS 模型可以获得文本中的对象类别、<对象,情感>以及<对象,方面,情感>, 而 JST 模型无法获得这些结果, 仅仅能获得方面类别、文本情感类别以及<方面,情感>。因此, 本文所给出的方法更适合于<对象,情感>、<方面,情感>以及<对象,方面,情感>联合识别, 也体现了本模型层次性的优势。

(3) JOAS 模型与 SVM 比较结果

SVM 分类器分类效果均低于 JOAS 模型。主要原因是 SVM 分类器忽略了对象、方面、情感之间的相互联系, 而在实际语料中, 情感词与所描述的方面和对象之间存在某种联系, 同样, 方面词与对象之间也存在某种联系, 具体实例见第 5.5 节。因此, 将对象、方面、情感分类视为独立无关的分类任务使 SVM 分类器的性能受到影响。

5.5 基于 JOAS 模型的方面情感词获取

利用 JOAS 模型 (参数选用 5.4 节最好的结果), 除了可以获得对象类别、方面类别、情感类别、<对象,情感>、<方面,情感>以及<对象,方面,情感>, 还可以获得特定领域下的方面情感词, 以概率由大到小排名前 40 的词中列举了部分方面和情感词汇, 如表 3 所示。

表 3 方面与情感对应的词汇表

方面	极性	词汇
空间	正	空间, 大, 后备箱, 后排, 可以, 满意, 不错, 足够, 东西, 够用, 很大, 宽敞, 座椅
	负	空间, 后排, 坐, 小, 放, 头部, 不能, 局促, 地方, 座椅, 前排, 储物空间
动力	正	动力, 档, 发动机, 油门, 快, 起步, 加速, 速度, 超车, 高速, 不错, 提速, 推背感
	负	起步, 肉, 动力, 开, 不, 档, 慢, 空调, 油门, 踩, 重, 次, 弱, 没有, 小
油耗	正	油耗, 公里, 跑, 油, 满意, 市区, 高速, 平均, 磨合期, 省油, 降, 低, 可以, 不错
	负	油耗, 不, 高, 高速, 公里, 大
舒适性	正	座椅, 不错, 舒服, 舒适性, 包裹性, 坐, 好, 很好, 感觉, 可以, 舒适, 满意, 真皮
	负	感觉, 胎噪, 座椅, 硬, 颠簸, 后排, 噪音, 舒适性, 减速, 明显, 路, 悬挂, 路面
内饰	正	内饰, 感觉, 中控, 不错, 软, 看, 喜欢, 设计, 做工, 好, 座椅, 中控台, 用料
	负	内饰, 塑料, 不, 硬, 中控, 配, 做工, 中控台, 没有, 塑料感, 地方, 不是, 座椅

由表 3 可以看出: 方面情感词可以获得在特定方面下的情感词, 例如, “空间”方面, 在正类中有“大”、“宽敞”、“足够”等, 负类中有“小”、“局促”等; “动力”方面, 在正类有“加速”、“超车”、“快”、“推背感”等, 负类中有“肉”、“弱”等等。对于获得的词汇, 可以用于构建特定领域情感词典和情感分类等任务。

6 结束语

本文在 LDA 的基础上, 拓展了五层结构, 给出了一个能同时挖掘商品评论数据中对象、方面及情感的生成式模型-JOAS 模型。模型能够将某类商品数据结构化, 同时可以获取其中的品牌、方面和情感信息, 从而指导消费者和生产厂家。在汽车评论数据上进行了多组实验, 实验表明, 本模型在方面分类和情感分类任务中, 分类能力与对比方法相当, 而且能够同时获取对象信息, 并取得较高的精确度。

本模型是对每个词采样情感标签, 对于否定句、转折等复合句不能很好的刻画, 在今后的工作中将本模型与深度学习结合, 构建出更加高效的对象、方面以及情感的联合模型。

参考文献

- [1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.
- [2] Lin C, He Y. Joint sentiment/topic model for sentiment analysis[C]// ACM Conference on Information and Knowledge Management. ACM, 2009:375-384.
- [3] He Y, Lin C, Alani H. Automatically extracting polarity-bearing topics for cross-domain sentiment classification.[C]// The Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, Usa. 2011:123-131.
- [4] Lin C, He Y, Everson R, et al. Weakly supervised joint sentiment-topic detection from text[J]. Knowledge & Data Engineering IEEE Transactions on, 2012, 24(6):1134-1145.
- [5] Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis[C]// Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February. 2011:815-824.
- [6] 孙艳, 周学广, 付伟. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报:自然科学版, 2013, 49(1):102-108.
- [7] Li F, Huang M, Zhu X. Sentiment analysis with global topics and local dependency.[C]// Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, Usa, July. 2010.
- [8] Moghaddam S, Ester M. On the design of LDA models for aspect-based opinion mining[C]// ACM International Conference on Information and Knowledge Management. 2012:803-812.
- [9] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]// International Conference on World Wide Web. ACM, 2007:171--180.
- [10] Titov, Ivan, McDonald, et al. A joint model of text and aspect ratings for sentiment summarization[J]. PROC. ACL-08: HLT, 2008:308--316.
- [11] Titov I, McDonald R. Modeling online reviews with multi-grain topic models[C]//Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 111-120.
- [12] Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, Mit Stata Center, Massachusetts, Usa, A Meeting of Sigdat, A Special Interest Group of the ACL. 2010:56-65.
- [13] Moghaddam, Samaneh, Ester, Martin. The FLDA model for aspect-based opinion mining: addressing the cold start problem[C]// International Conference on World Wide Web. 2013:909-918.
- [14] Rao Y, Li Q, Liu W, et al. Affective topic model for social emotion detection[J]. Neural Networks the Official Journal of the International Neural Network Society, 2014, 58(5):29-37.
- [15] Chen Z, Mukherjee A, Liu B. Aspect extraction with automated prior knowledge learning[C]// ACL. 2014.
- [16] Wu Y, Ester M. FLAME: A probabilistic model combining aspect based opinion mining and collaborative

- filtering[C]// Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015:199-208.
- [17] Mcauley J, Pandey R, Leskovec J. Inferring networks of substitutable and complementary products[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015:785-794.
- [18] Samha A K, Li Y, Zhang J. Aspect-based opinion mining from product reviews using conditional random fields[J]. 2015.
- [19] Walsh B. Markov chain monte carlo and gibbs sampling. WWW document[J]. Notes, 2004:497-537.
- [20] 冯淑芳, 王素格. 面向观点挖掘的汽车评价本体知识库的构建[J]. 计算机应用与软件, 2011, 28(5):45-47.



李大宇 (1992—), 男, 硕士生, 主要研究方向为情感分析;
E-mail: 949007264@qq.com



李旻 (1988—), 女, 博士生, 主要研究方向为情感分析。
E-mail:



王素格 (1964—), 女, 教授, 博士, 主要研究方向为自然语言处理、情感分析。
E-mail: wsg@sxu.edu.cn