

# 基于主题情感联合分析的游客画像研究

李 琴<sup>1</sup>, 李少波<sup>2</sup>, 胡 杰<sup>1</sup>

(1. 贵州财经大学 大数据统计学院, 贵阳 550000; 2. 贵州大学 机械工程学院, 贵阳 550000)

**摘 要:** 网络文本作为现代游客承载感知和表达观点的载体, 已成为游客画像构建与分析的重要数据来源。现有的自然语言处理技术在游客画像的挖掘过程中主要关注游客的需求和情感, 缺少技术与旅游应用的有效衔接, 然而现有的文本挖掘技术中文本的主题和情感通常被割裂分析, 缺乏相互指向性, 无法有效提取用户细粒度的意见。提出一种基于变分自编码的有监督主题情感联合分析模型。将词频权重引入到先验知识中, 同时通过截断高斯模型构造变参数, 有效捕获离散数据中的相关性, 利用情感标签辅助主题的训练和生成, 以提升主题挖掘及情感预测的准确率。通过变分自编码模型计算贝叶斯主题模型的后验分布, 采用主题分布下的情感分类预测实现主题情感的联合分析。实验结果表明, 当主题数为 10~100 时, 该模型的情感预测平均准确率约为 85%, 相比 LDA、SAGE、NVDM 模型, 能够有效挖掘酒店用户评论的特征。

**关键词:** 游客画像; 变分自编码器; 主题情感联合分析; 意见挖掘; 隐含狄利克雷分布模型

开放科学(资源服务)标志码(OSID):



中文引用格式: 李琴, 李少波, 胡杰. 基于主题情感联合分析的游客画像研究[J]. 计算机工程, 2022, 48(6): 278-287, 294.

英文引用格式: LI Q, LI S B, HU J. Research on tourist portrait based on joint topic-sentiment analysis[J]. Computer Engineering, 2022, 48(6): 278-287, 294.

## Research on Tourist Portrait Based on Joint Topic-Sentiment Analysis

LI Qin<sup>1</sup>, LI Shaobo<sup>2</sup>, HU Jie<sup>1</sup>

(1. College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang, 550000, China;

2. School of Mechanical Engineering, Guizhou University, Guiyang, 550000, China)

**[Abstract]** As the carrier of modern tourists' perception and expression of views, network text has become an important data source for the construction and analysis of tourist portrait. The existing natural language processing technology focuses on the needs and emotions of tourist portraits, and lacks an effective connection between technology and tourism applications. However, in the existing text mining technology, the topic and sentiment of text are usually separated and analyzed, show a lack of mutual directivity, and cannot effectively extract users' fine-grained opinions. A supervised joint topic-sentiment analysis model based on Variational Auto-Encoders (AVEs), is proposed. The word frequency weight is introduced into the prior knowledge, and the variable parameters are constructed by Gaussian Stick-Breaking model to effectively capture the correlation in the discrete data. The sentiment label is used to assist the topic training and generation, to improve the accuracy of topic mining and emotion prediction. The posterior distribution of the Bayesian topic model is calculated using the AVEs model, and the sentiment classification prediction under topic distribution is used to realize the joint topic-sentiment analysis. The experimental results show that the average accuracy of this model is about 85% when the number of topics is 10~100. Compared with LDA, SAGE and NVDM models, this model can effectively mine the characteristics of hotel user comments.

**[Key words]** tourist portrait; Variational Auto-Encoders (VAEs); joint topic-sentiment analysis; opinion mining; Latent Dirichlet Allocation (LDA) model

DOI: 10.19678/j.issn.1000-3428.0061490

## 0 概述

游客画像是对游客属性标签化的过程, 主要应

用于旅游目的地的精准营销、个性化服务、游客行为分析、舆情治理等方面, 是实现智能化旅游的关键。酒店作为旅游经济过程中的重要因素, 与旅游经济

**基金项目:** 贵州省科技计划项目(黔科合基础-ZK[2021]337); 贵州省教育厅青年科技人才成长项目(黔教合KY字[2021]141); 贵州财经大学引进人才科研启动项目(2021YJ003)。

**作者简介:** 李 琴(1985—), 女, 讲师、博士, 主研方向为自然语言处理、旅游大数据分析; 李少波, 教授、博士; 胡 杰, 讲师、博士。

**收稿日期:** 2021-04-27 **修回日期:** 2021-07-19 **E-mail:** qinlee85@126.com

起着相互促进的作用。酒店是否符合现阶段游客的需求成为游客衡量旅游目的地的重要因素之一。通过挖掘分析获取不同群体的需求或喜好特点继而推荐符合不同群体需求的酒店,是提升游客体验和酒店运营的有效手段。现代旅游过程以社会互动和旅游信息交换为特征,其产物——游客生成文本,能够反映游客的喜好、感知和需求信息,通过对游客生成文本进行分析,准确获取不同群体的情感喜好等信息对旅游酒店推荐具有重要意义。

随着深度学习的发展,以文本数据为主的自然语言处理技术异军突起<sup>[1]</sup>,在主题挖掘和情感分析等领域取得重大进展<sup>[2]</sup>。在大多数情况下,文本的主题和情感仍被割裂开来分析,然而在实际情况中,通常要求主题和情感具有相互指向性,例如“美丽的”指向具体对象如“花园”或“花园”指向情感要素如“美丽的”。如何进行主题和情感的联合分析成为研究热点。大量无监督主题情感分析模型应运而生,如JST模型<sup>[3]</sup>、ASUM模型<sup>[4]</sup>、JMTS<sup>[5]</sup>等。这类无监督主题情感模型认为词的生成与主题和情感都相关,通过对每个句子或每个词进行情感标签和主题标签采样,以生成句子的主题和情感对。另一类无监督主题情感模型(如WLDA<sup>[6]</sup>、TSLDA<sup>[7]</sup>、JST-RR<sup>[8]</sup>等)通过引入先验知识(如互信息、主观性词典、主题意见词对、文本情感等),在获取主题的同时提升情感检测率。这类模型并不是完全无监督,其利用先验知识诱导先验,从而增强主题模型的稀疏性。这2类模型均以隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)为基础模型,具有较强的挖掘能力。LDA作为一种贝叶斯生成模型,主要依赖于关键词词频信息。但是LDA模型缺乏先验信息的指导且仅适用于长文本的分析,采用吉布斯采样进行后验分布计算,当文档数量多而主题个数较少时,LDA模型的训练速度相对较慢,并且需要在数学上重新推导新的推理算法进行更改。LDA作为一种无监督模型,缺乏标签的约束,其训练得到的主题通常表达解释能力较差。有监督学习利用标签数据的正向回馈,其准确率优于无监督学习<sup>[9]</sup>。SLDA模型<sup>[10]</sup>将元数据作为标签(如情感评分等),以辅助推断和预测标签相关的主题,相对无监督的LDA,该模型具有更优的预测能力。

LDA作为概率主题模型中简单且经典的模型,为主题模型提供了一个标准框架,在学术界和工业界具有广泛的研究和应用价值,但其自身的局限性却不容忽视。随着变分自编码(Variational Auto-Encoders, VAEs)模型的提出,使用变分自编码深度学习在特征提取方面(如情感、主题等)取得巨大的成功<sup>[11-12]</sup>。VAEs是一种深度生成模型,又称为AEVB算法,该模型基于变分的贝叶斯理论,将编码器和解码器设置为神经网络,通过迭代优化过程学习最佳的编码-解码方案。结果表明<sup>[13]</sup>,相比使用吉

布斯采样的LDA主题模型,VAEs在主题模型上的应用能够有效挖掘主题,且更易扩展。此外,重参数化技巧RT(Reparameterization Trick)及SGVB估计算法建立AEVB算法的梯度反向传播机制。RT在技术方面的提高使得更多的分布能被应用在VAEs中,同时为VAEs近似复杂概率模型提供更多的可能性。

研究工作表明<sup>[14-15]</sup>,先验分布的复杂度及超参数的选择对于深度生成模型或贝叶斯神经网络的性能具有重要意义。本文提出基于变分自编码的有监督主题情感联合分析模型SJST-VAE。通过先验知识和情感标签辅助主题的训练和生成,利用截断高斯模型变分参数构造适合主题挖掘过程的神经变分推断形式,采用主题分布下的情感分类预测实现主题情感的联合分析。

## 1 相关工作

变分自编码在主题概率模型中得到广泛应用<sup>[16]</sup>。LDA的任何变体都需派生自定义推理算法,然而变分自编码具有较强适应数据特征的能力,其推理方法为隐藏变量建模提供强大的架构,具有更强的可扩展性。AVITM(Autoencoding Variational Inference for Topic Models)模型<sup>[13]</sup>通过构建变分自编码与主题模型的桥梁,降低Dirichlet先验和组件坍塌(类似于先验信任的局部最优)对AEVB算法产生的影响。针对传统主题模型在短文本上表现较差的问题,文献[17]利用词向量和主题向量的点积构建词的主题分布,并定义了词的上下文表征以区分一词多义的现象,提出一种利用词向量语义关系辅助主题挖掘的嵌套变分贝叶斯的主题模型。文献[18]提出使用Gumbel-Softmax模型和高斯混合模型建模变分自编码主题类别分布,解决局部最优的问题,并分析选择不同分布模型对主题生成的影响。文献[19]利用动态因子图模拟主题在时间上的动态变化,基于变分自编码构建动态的主题模型。针对传统主题模型无法动态确定主题数量的问题,文献[20]基于自编码变分推断的架构,提出一种循环神经主题模型,以发现从概念上无界限的主题。这些模型虽然根据变分自编码易扩展的特性展现出在主题挖掘上的优势,但是缺乏主题的情感指导或主题与情感的联合分析。

主题情感的联合分析是一种细粒度的意见挖掘,其目标是从文本主观评论中获取情感倾向的观点或情感要素。在旅游领域中具体的实际应用尤其是旅游推荐具有重要意义。结合深度学习的思想,基于方面或目标实体的情感分类(TABSA)虽然在挖掘文本特征信息和对应情感属性上取得较大进展,但是大多依赖于文本类别、特征属性及对应情感类别的标注,使得实际工作面临较大的困难。传统的无监督主题情感模型在一定程度上解决数据标注

缺乏的问题,但因计算复杂度高且时间消耗久等问题,导致模型难以扩展。变分自编码主题模型的实现成为解决该问题的关键。文献[18]基于传统情感主题联合模型 JST,预先定义特征种子词,通过对 AVITM 模型进行扩展,实现变分自编码的无监督情感与主题的分析。但是任何无监督的模型都无法假设现实中的所有情况<sup>[10]</sup>,其相较于有监督模型的准确率较低。根据当前旅游社交网站中文本评论及情感评分易于获取的特点,本文基于 AVITM 模型,以逻辑或知识表示先验知识,利用情感监督主题识别辅助预测情感分类,从而实现主题情感的联合分析和酒店游客的特征画像。

## 2 LDA 模型

LDA 模型由 Dirichlet 先验的主题分布得名, Dirichlet 先验的选择对于可解释性主题的获得具有重要作用。在 LDA 模型中,主题被看作相关主题的词汇分布,每个文档被看作多个主题分布。为生成文档  $d$ ,该过程会随机选择主题分布  $\theta_d$ ,从主题分布中随机选择一个主题  $z_{d,n}$ ,从相应主题或主题的词汇分布  $\beta_k$  随机选择一个词来生成文档中每个可观测词  $w_{d,n}$ 。因此, LDA 中文档  $d$  的生成如下:

$$p(w|\alpha, \beta) = \int_{\theta} \left( \prod_{n=1}^N \sum_{z_n=1}^k p(w_n|z_n, \beta) p(z_n|\theta) \right) p(\theta|\alpha) d\theta \quad (1)$$

$$z_n \sim \text{Multinomial}(1, \theta) \quad (2)$$

$$w_n \sim \text{Multinomial}(1, \beta_{z_n}) \quad (3)$$

其中:  $\alpha$  为 Dirichlet 分布的超参数。在多项分布假设下,  $\theta$  和  $\beta$  之间的耦合导致隐变量  $\theta$  和  $z$  后验分布的推理难以计算,需要借助各种近似方法。LDA 模型采用吉布斯采样方法,即一类 MCMC (Markov Chain Monte Carlo) 算法,通过抽取大量样本估计真实的后验分布,但该方法计算复杂度高且时间消耗量大。

## 3 变分自编码框架下主题分布的参数化

针对 LDA 模型中后验分布难以计算的问题,研究人员提出变分推理方法,通过优化过程寻求一种变分分布近似真实的后验分布。MFVI (Mean-Field Variational Inference) 方法是一种比较经典的变分推理方法,但是由于计算原因难以扩展到新的模型。AEVB 算法旨在以一种“黑匣子”推理方法来解决该问题,该算法利用推断和学习使得简单的采样就能进行有效的近似推断,不需要复杂的迭代推理方式(如 MCMC)。

在 LDA 模型中,隐变量  $z$  是离散变量,无法进行重参数化处理,通过求和运算折叠  $z$  变量,即将式(1)转变为:

$$p(w|\alpha, \beta) = \int_{\theta} \left( \prod_{n=1}^N p(w_n|\beta, \theta) \right) p(\theta|\alpha) d\theta \quad (4)$$

其中:  $w_n|\beta, \theta \sim \text{Multinomial}(1, \beta\theta)$ 。

因此,后验分布难以计算的问题转化为评估  $\theta$  和  $\beta$  的分布。VAE 使用自编码学习  $\theta$  和  $\beta$  的分布,同时通过拉普拉斯将原始 Dirichlet 先验分布近似为变分分布。在 LDA 主题模型中,主题变量的先验  $\theta = (\theta_1, \theta_2, \dots, \theta_K)$  ( $K$  为主题个数) 被定义为 Dirichlet 分布,经过变分推理,主题变量的 Dirichlet 先验  $p(\theta|\alpha)$  通过拉普拉斯被近似为一个多元高斯分布。多元高斯分布由均值向量  $\mu$  和对角协方差矩阵  $\Sigma$  定义,其中

$$\mu = (\mu_1, \mu_2, \dots, \mu_K), \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_K^2 \end{bmatrix}, \text{ 所以 } p(\theta|\alpha)$$

近似为  $q(\theta) = \mathcal{LN}(\theta|\mu, \Sigma)$ , 其中  $\mathcal{LN}$  是逻辑正态分布。逻辑正态分布更能促进主题一致性。

通过拉普拉斯近似计算得到多元高斯分布的均值向量  $\mu$  和对角协方差矩阵  $\Sigma$ , 如式(5)和式(6)所示:

$$\mu_k = \ln \alpha_k - \frac{1}{K} \sum_i \ln \alpha_i \quad (5)$$

$$\Sigma_{kk} = \frac{1}{\alpha_k} \left( 1 - \frac{2}{K} \right) + \frac{1}{K^2} \sum_i \frac{1}{\alpha_i} \quad (6)$$

在变分自编码框架下,将观测数据文档  $w$  词序列作为输入,将 2 个推断网络作为前向神经网络  $f_{\mu}(w, \delta)$  和  $f_{\Sigma}(w, \delta)$ , 其中,  $\delta$  为推断网络的参数,即变分参数,  $\delta = (\bar{\mu}, \bar{\Sigma})$ , 从而估计  $\bar{\mu}$  和  $\bar{\Sigma}$  的值,每个网络的输出均为  $K$  维向量。

变分自编码架构使用 RT 操作进行反向传播,令  $x = \bar{\mu} + \epsilon \bar{\Sigma}^{\frac{1}{2}}$ , 通过采样  $\epsilon$  生成样本  $x$ , 其中  $x \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ ,  $\epsilon \sim \mathcal{N}(0, 1)$ 。

变分分布的构造有多种形式,高斯分布是其经典的变分分布形式。在变量原始分布未知的情况下,高斯分布可以为噪声和不确定性建模。通过高斯先验分布和 RT 技术为变分分布建立无偏差或低方差的梯度估计器,如 SGVB。

### 3.1 单高斯分布模型

设  $x \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ , 隐变量  $\theta$  如式(7)所示:

$$\theta = \text{soft max}(W_1^T x) = \text{soft max}\left(\bar{\mu} + \epsilon \bar{\Sigma}^{\frac{1}{2}}\right) \quad (7)$$

$\theta$  服从单高斯分布模型 (Gaussian SoftMax, GSM), 即:

$$\theta \sim G_{\text{GSM}}(\bar{\mu}, \bar{\Sigma})$$

其中:  $W_1$  为线性变换权重,偏差项做了省略处理。

### 3.2 截断高斯模型

SB(Stick-Breaking)过程被用于主题变量 Dirichlet 过程的建设性定义,为其先验提供初始关联权重。在截断高斯(GSB)模型构建过程中,通过逐次分割单位为1的区间顺序获取高斯先验,其中 $\theta_k$ 表示每个分量。SB构建过程如图1所示。

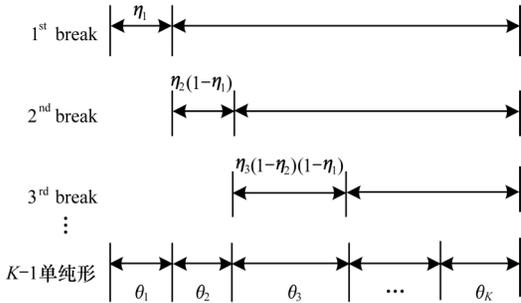


图1 SB构建过程

Fig.1 SB construction process

设第1个类别的概率为分割比例 $\eta_1$ ,其余比例 $1 - \eta_1$ 为后续的分割计算。高斯先验的每一维计算如式(8)所示:

$$\theta_k = \begin{cases} \eta_k \prod_{j < k} (1 - \eta_j), & k < K \\ \prod_{j \leq K} (1 - \eta_j), & k = K \end{cases} \quad (8)$$

不同的 $K$ 值需满足 $\sum_{k=1}^K \theta_k = 1$ 。多项式概率参数的建模被转化为二项式概率参数的对数建模。

设高斯样本 $\mathbf{x} \in \mathbb{R}^K$ ,  $\mathbf{W}_2 \in \mathbb{R}^{K \times (K-1)}$ , 则 $\theta \sim G_{\text{GSB}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ , 其构造过程如式(9)所示:

$$\mathbf{x} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}), \boldsymbol{\eta} = \text{sigmoid}(\mathbf{W}_2^\top \mathbf{x}), \boldsymbol{\theta} = f_{\text{SB}}(\boldsymbol{\eta}) \quad (9)$$

与Dirichlet过程的SB定义相比,高斯过程的SB为神经变分推断提供更合适的形式。在高斯先验的分配过程中,SB构建过程缺乏控制力,SB先验可以更好地保留类别边界,为半监督学习提供有效的正则化。同时,SB过程能够降低模型对主题数量变化的敏感度,更具稳定性。

### 4 本文模型

无监督主题模型缺乏有效先验知识的指导或监督学习中数据的标注。针对该问题,本文基于LDA模型,引入词频相对主题的权重,以影响主题的先验分布,从而指导主题的生成,同时通过情感标签监督生成主题,利用主题特征预测情感分类,从而实现主题情感的联合分析。本文模型SJST-VAE以VAE为主要架构,主要由先验知识的指导、情感标签的监督、变分目标损失函数的计算和主题情感的联合分析这4个部分组成。SJST-VAE模型架构如图2所示。

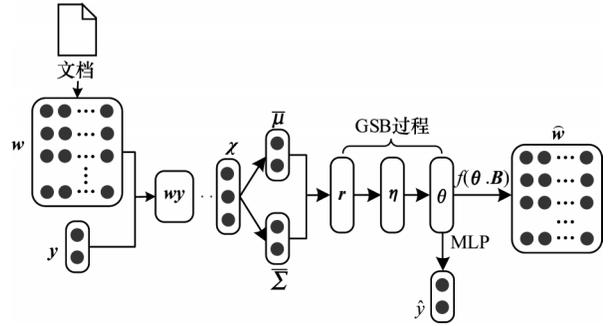


图2 SJST-VAE模型架构

Fig.2 Framework of SJST-VAE model

#### 4.1 先验知识的指导

VAE架构的优势是为编码网络提供一种可以引入先验信息的扩展方法。本文模型SJST-VAE以词的Bag-of-words表征作为输入,通过变分自编码网络获取文档的特征。此外,针对文本中权重过高的词大多不能进行局部主题表示的问题,如IMDB语料库中,单词“film”或“movie”在主题模型学习中往往相对不重要,本文通过弱化词频过高的词,设置背景术语从而获取相对常见的词,以此促进主题的一致性。

假设语料由 $D$ 个文档组成,语料词典大小为 $V$ ,如图2所示,在变分自编码网络框架下,SJST-VAE模型以文档词序列 $\mathbf{w}$ 作为输入,通过2个MLP推断网络 $f_{\bar{\boldsymbol{\mu}}}(\mathbf{w})$ 和 $f_{\bar{\boldsymbol{\Sigma}}}(\mathbf{w})$ 变分近似为具有对角正态先验的文本表征 $\mathbf{r}$ ,即 $\mathbf{r} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ ,如式(10)所示:

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= f_{\bar{\boldsymbol{\mu}}}(\mathbf{w}) = \mathbf{W}_{\bar{\boldsymbol{\mu}}} \mathbf{w} + \mathbf{b}_{\bar{\boldsymbol{\mu}}} \\ \bar{\boldsymbol{\Sigma}} &= f_{\bar{\boldsymbol{\Sigma}}}(\mathbf{w}) = \mathbf{W}_{\bar{\boldsymbol{\Sigma}}} \mathbf{w} + \mathbf{b}_{\bar{\boldsymbol{\Sigma}}} \end{aligned} \quad (10)$$

从而获得 $\mathbf{r}$ 的近似后验分布,如式(11)所示:

$$q_{\delta}(\mathbf{r}|\mathbf{w}) = \mathcal{N}(\mathbf{r}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \quad (11)$$

其中: $\mathbf{W}_{\bar{\boldsymbol{\mu}}}$ 、 $\mathbf{b}_{\bar{\boldsymbol{\mu}}}$ 、 $\mathbf{W}_{\bar{\boldsymbol{\Sigma}}}$ 、 $\mathbf{b}_{\bar{\boldsymbol{\Sigma}}}$ 分别为2个神经网络的线性参数,且 $\delta = (\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ 。为解决神经网络反向传播的问题,通过从高斯噪声中采样 $\epsilon$ ,且 $\epsilon \sim \mathcal{N}(0, 1)$ ,使得 $\mathbf{r} = \bar{\boldsymbol{\mu}} + \epsilon \bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}$ 。通过GSB构造将文本表征 $\mathbf{r}$ 转化为单纯形 $\boldsymbol{\theta}$ ,其映射过程如图3所示。

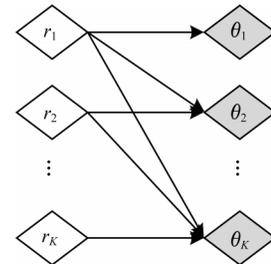


图3 SB映射过程

Fig.3 SB mapping process

单纯形 $\boldsymbol{\theta}$ 如式(12)所示:

$$\boldsymbol{\theta} \sim G_{\text{GSB}}(\mathbf{r}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \quad (12)$$

本文设置稀疏诱导先验,即正态指数复合先验,置于权重矩阵 $\mathbf{B} \in \mathbb{R}_{K \times V}$ (主题词分布矩阵的初始化),使模型学习到词频相对主题的权重信息,同时定义

背景术语  $\mathbf{d} \in R_v$ , 表示词频中所有词的词频对数值, 旨在通过  $\mathbf{B}$  与  $\mathbf{d}$  的偏离程度将主题权重倾向于文档中出现频率大致相同的常见词, 而不是词频过高的词。权重矩阵  $\mathbf{B}$  的正态指数复合先验过程如式(13)和式(14)所示:

$$\tau_{k,v} \sim \text{Exponential}(\zeta) \quad (13)$$

$$B_{k,v} \sim \mathcal{N}(0, \tau_{k,v}) \quad (14)$$

其中:  $\zeta > 0$  为指数分布率参数。重构文档为:

$$\hat{\mathbf{w}} = \mathbf{d} + \boldsymbol{\theta}^T \mathbf{B} \quad (15)$$

整个神经网络的损失函数如式(16)所示:

$$\mathcal{L}(\mathbf{w}, \hat{\mathbf{w}}) = \mathbf{w} - \hat{\mathbf{w}}^2 \quad (16)$$

#### 4.2 情感标签的监督

除了能有效推断文本主题外, 本文模型 SJST-VAE 还能推断文本的潜在表达及预测文本的情感倾向, 利用情感标签对主题生成前后进行监督。生成主题后的监督是在可观测词的前提下, 完成主题模型的变分自编码解码的过程, 利用多层神经网络进行训练, 将预测标签与真实标签的交叉熵作为损失函数, 从而实现情感标签的预测。情感预测标签的计算如式(17)所示:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y} | f_{\mathbf{y}}(\boldsymbol{\theta})) \quad (17)$$

其中:  $f_{\mathbf{y}}$  为多层神经网络。

在主题生成过程中, 情感标签还用于监督主题训练过程, 以促进局部主题的生成。SJST-VAE 模型在预测情感标签时, 利用 one-hot 编码表征文档的情感标签  $\mathbf{e}_y$ , 并对编码器网络进行训练, 情感标签  $\mathbf{e}_y$  被用于构建文本表征的特征, 如(18)和式(19)所示:

$$q_{\delta}(\mathbf{r} | \chi(\mathbf{w}, \mathbf{e}_y)) = \mathcal{N}(\mathbf{r} | \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \quad (18)$$

$$\chi(\mathbf{w}, \mathbf{e}_y) = f_g([\mathbf{W}_x \mathbf{w}; \mathbf{W}_y \mathbf{e}_y]) \quad (19)$$

其中:  $f_g$  为多层感知器;  $\mathbf{W}_x$  和  $\mathbf{W}_y$  为线性权重参数。

在训练过程中, 情感类别标签作为可观测变量。在测试时, 本文模型考虑所有可能的情感标签向量, 如正向或负向, 使得文档中所有词概率对数和最大的标签为所预测标签, 如式(20)所示:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^V \ln p(v_i | \mathbf{r}, \mathbf{e}_y) \quad (20)$$

其中:  $v_i$  为词典中的词,  $i = \{1, 2, \dots, V\}$ 。

#### 4.3 变分目标损失函数

与传统变分推理类似, SJST-VAE 模型构造一个 Dirichlet 先验的拉普拉斯近似, 使 Dirichlet 分布可以近似为逻辑正态分布。本文假设 Dirichlet 先验是对称的, 即所有超参数  $\alpha$  取相同值, 由式(5)和式(6)可得:

$$\mu_k = 0, \Sigma_{kk} = \frac{K-1}{\alpha K} \quad (21)$$

在变分自编码架构下, 本文设计高斯变分分布, 以近似后验分布  $q_{\delta}(\mathbf{r} | \mathbf{w}, \mathbf{e}_y)$ 。模型学习的目标是使近似后验分布尽可能接近于真实后验分布  $p(\mathbf{r} | \alpha)$ 。本

文采用 KL 散度进行相似度计算, 找到能使 KL 散度尽可能小的变分参数  $\delta = (\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ , 如式(22)所示:

$$q_{\delta}^*(\mathbf{r} | \chi(\mathbf{w}, \mathbf{e}_y)) = \underset{\delta}{\operatorname{argmin}} D_{\text{KL}}(q_{\delta}(\mathbf{r} | \chi(\mathbf{w}, \mathbf{e}_y)) \| p(\mathbf{r} | \alpha)) \quad (22)$$

通过一系列计算推演, 式(22)转换为使变分下界 ELBO 最大化, 其变分下界如式(23)所示:

$$\mathcal{L}(\mathbf{w}) \approx \ln p(\mathbf{w} | \mathbf{r}) + \ln p(\mathbf{y} | \mathbf{r}) - D_{\text{KL}}(q_{\delta}(\mathbf{r} | \chi(\mathbf{w}, \mathbf{e}_y)) \| p(\mathbf{r} | \alpha)) \quad (23)$$

其中: KL 散度为正则项; 其他部分为重构损失。KL 散度如式(24)所示:

$$D_{\text{KL}}(q_{\delta}(\mathbf{r} | \chi(\mathbf{w}, \mathbf{e}_y)) \| p(\mathbf{r} | \alpha)) = \frac{1}{2} \left( \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{\Sigma}}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}) - K + \ln \frac{|\boldsymbol{\Sigma}|}{|\bar{\boldsymbol{\Sigma}}|} \right) \quad (24)$$

#### 4.4 主题情感的联合分析

SJST-VAE 模型通过变分自编码网络获取各文档的主题分布  $\theta_d, \theta_{d_1}, \dots, \theta_{d_n}$ , 并将其作为输入, 通过文档的情感监督对 MLP 神经网络进行训练, 从而实现情感分类的预测。

本文假设文档数据集有  $K$  个主题, 并将只包含第  $k$  个主题的文本主题分布  $t_k$  定义为除第  $k$  个分量为 1 外, 其余各分量均为 0 的向量。因此, 为获得第  $k$  个主题的情感分布, 模型以  $t_k$  作为输入向量进行情感预测, 其中  $k = 1, 2, \dots, K$ , 以获取各个主题下的情感概率分布。

### 5 实验结果与分析

#### 5.1 数据集与参数设置

本文将 IMDB 语料集作为评估 SJST-VAE 模型的数据集, 该数据集包含 50 000 条电影评论, 其中 25 000 条负面评论和 25 000 条正面评论, 且训练集 25 000 条和测试集 25 000 条。在数据集中所有单词通过预处理均被转化为字母小写形式, 并删除了标点符号、数字及小于 3 个字符和停用词表中的所有单词。词典由在大多数文档中都出现的单词组成, 大小设为 2 000。在模型训练过程中, 本文使用 softplus 激励函数、Adam 优化器(参数设为 0.99), 学习率设为 0.002, 批量大小设为 200,  $\epsilon$  采样数量设为 1, 训练迭代次数设为 200。在测试文档估计 ELBO 值时,  $\epsilon$  采样数量设为 20。

#### 5.2 评估标准

SJST-VAE 模型是基于情感监督进行主题情感联合分析。研究人员给出不同的衡量标准, 如困惑度、相关性、稀疏度等, 以客观评价主题获取的优劣程度。困惑度表示文档属于哪个主题的不确定性, 困惑度越低, 聚类效果越好, 主题与主题的区别性越

强;相关性表示模型获取主题的 top- $n$  个词的语义一致性,一致性越高表示主题可解释性越好;稀疏度在一定意义上表示模型的可解释性,因为每个主题能够接受被描绘的词往往是有限的,主题的词分布矩阵越稀疏(即稀疏度越大),可解释性越强。虽然这3种评估标准具有一定的有效性,但其不能完全作为评估标准,有时需要直观的主题表示进行评估。模型主题困惑度如式(25)所示:

$$\text{perplexity}(C) = \exp \left( - \frac{\sum_{d=1}^D \ln p(w_d)}{\sum_{d=1}^D N_d} \right) \quad (25)$$

其中: $C$ 为测试语料且包含 $D$ 篇文档; $N_d$ 为每篇文档包含词的数量; $p(w_d)$ 为文档 $d$ 中词产生的概率。

本文采用 NPMI(Normalized Pointwise Mutual Information)对文本语料主题相关性进行评估。PMI 主要用于度量一些词的共现,以此判定词的相关性,PMI 如式(26)所示:

$$\text{PMI}(w_1, w_2) = \ln \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \quad (26)$$

对于 PMI 的正则化有多种选择,如通过  $-\ln p(w_1)$  和  $-\ln p(w_2)$  的乘积或通过  $-\ln p(w_1, w_2)$  正则化。本文以后者作为正则化选项,该正则化过程规范了上限和下限,具有较优的性能。因此,NPMI 如式(27)所示:

$$i_n(w_1, w_2) = \left( \ln \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \right) / (-\ln p(w_1, w_2)) \quad (27)$$

### 5.3 模型评估与结果分析

本文研究先验知识和情感监督对主题获取的影响,因此,评估分析了模型在相同实验条件和参数设置条件下有先验知识和无先验知识的性能对比,以及有情感监督和无情感监督的性能对比,并验证了在 GSM 和 GSB 构造下不同主题数目设置对主题分布性能的影响。为验证本文方法的有效性,本文将 SJST-VAE 模型与其他 3 种基准主题模型进行对比。这 3 种基准模型分别为 LDA、SAGE(Sparse Additive Generative Model)<sup>[19]</sup>、NVDM(Neural Variational Document Model)<sup>[20]</sup>。

IMDB 语料的平均主题分布如图 4 所示。从图 4 可以看出,在 GSM 过程中获得 100 个主题的平均主题分布情况大致相同,而在 GSB 过程中的平均主题分布值在接近第 10 个主题位置后逐渐递减,在大概第 20 个主题后递减速度尤为明显,直至第 40 个主题后几乎没有分布。这是因为 GSB 过程在建立混合模型时,其 SB 结构隐含地假定了主题的顺序,前一个主题获得足够的梯度来更新主题分布。同时,SB 结构的稀疏性使得尾部的主题被采样的可能性较小,模型对于超参数(主题数目)的变化会变得不太敏感,当主题设置数目远远超过模型需要的数目时,GSB 过程的稳定性更强,而且更加有利于主题数目的设置。

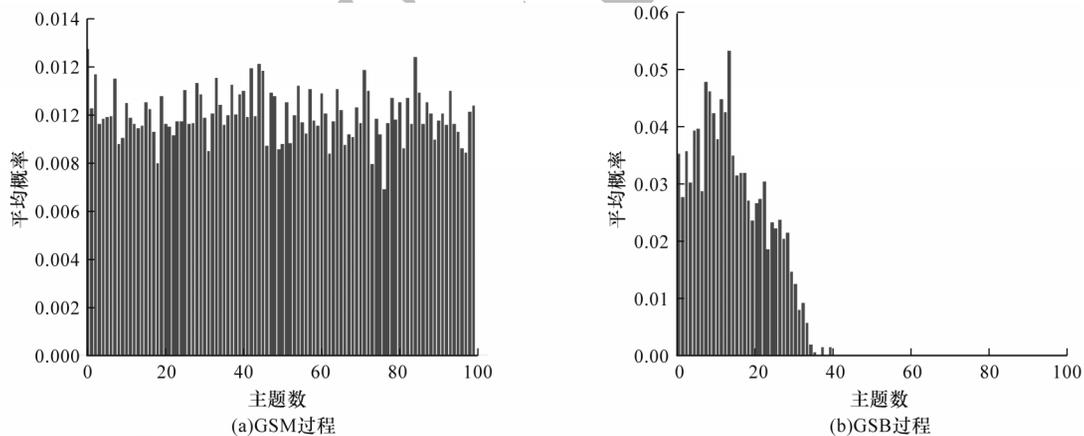


图4 IMDB 平均语料主题分布

Fig.4 Average topics distribution of IMDB corpus

主题数为 10~100 及 100~500 时随模型测试集困惑度的变化情况如图 5 所示。从图 5 可以看出,主题数从 10~100 的变化过程中,GSB 过程的主题困惑度略优于 GSM 过程,随着主题数从 100~500 逐渐增大,GSB 过程在主题困惑度上表现出的优势越来越明显,说明 GSB 过程不会因主题数目变化而发生大幅波动,验证了 GSB 过程的稳定性。

现有的主题模型主要在数据集 20newsgroups 进行训练,该数据集缺少情感标签数据。因此,本

文以 IMDB 数据集为对象,选择具有代表性的基准模型对主题挖掘性能进行评估。LDA 是经典的模型,几乎所有模型都以此为基础;SAGE 模型引入恒定背景分布的对数频率,以防止过度拟合,即通过稀疏诱导先验加强模型主题的稀疏性,具有较强的鲁棒性;NVDM 模型首次将神经变分框架的生成模型引入到文本建模中,旨在为每个文档提取一个连续的语义潜在变量,并应用于构建主题分类。

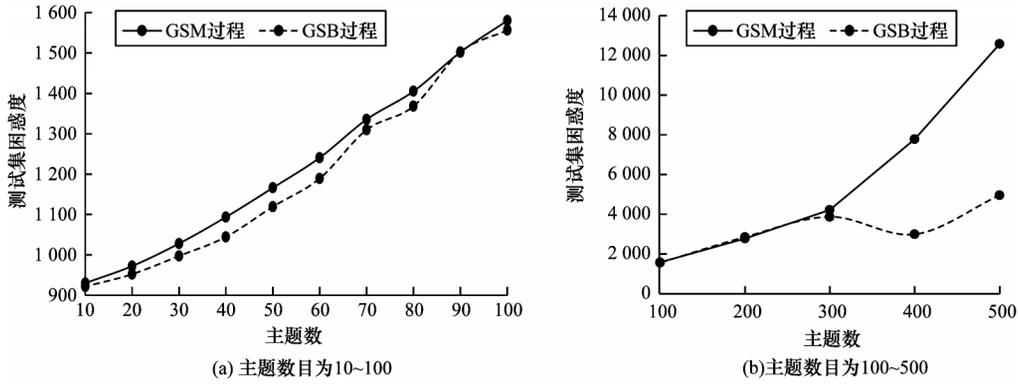


图5 主题困惑度随主题数的变化趋势

Fig.5 Trend of topic perplexities with number of topics

本文将主题数目设置为10和50,不同基准模型的主题困惑度、相关性和稀疏度的对比结果如表1所示。

表1 不同模型的主题评估指标

Table 1 Topic evaluation indicators of different models

主题模型	10个主题			50个主题		
	困惑度	相关性	稀疏性	困惑度	相关性	稀疏性
SJST-VAE(GSM)	930	0.138	0	1 166	0.134	0
SJST-VAE(GSB)	921	0.145	0.002	1 119	0.105	0.002
SJST-VAE-bg(GSB)	921	0.136	0	1 109	0.108	0
SJST-VAE-senti(GSB)	912	0.146	0	1 122	0.122	0
LDA	—	—	—	1 508	0.130	0
SAGE	—	—	—	1 767	0.120	0.790
NVDM	—	—	—	1 748	0.060	0

相比基准模型, SJST-VAE模型在主题困惑度上具有较强的优势; SJST-VAE模型的主题相关性低于LDA和SAGE模型,但优于同是变分自编码框架的NVDM模型; SAGE模型的主题稀疏性仍占有绝对优势, SJST-VAE模型相对于其他模型略有改进。同

时,本文对比SJST-VAE模型在背景术语缺失(SJST-VAE-bg)和情感监督缺失(SJST-VAE-senti)以及GSM和GSB过程下的主题性能评估。在主题数目设置为10和50时,GSB过程中SJST-VAE模型、SJST-VAE-bg模型、SJST-VAE-senti模型的主题困惑度均优于GSM过程中的SJST-VAE模型,说明GSB过程在主题困惑度的表现上具有绝对优势。当主题数目设置为50时, SJST-VAE模型的GSM过程的相关性具有一定优势。GSB过程的SJST-VAE模型相对于SJST-VAE-bg模型和SJST-VAE-senti模型的稀疏性略有提高,说明SJST-VAE模型具有较强的主题可解释性。由于高频背景术语的缺失以及情感词的加入使得模型在主题数目增多的情况下,发生主题一致性降低的情况。其原因为随着主题数目增多时,无明显意义主题词出现的概率会增大,而主题一致性的计算基于词的共现,高频词的缺失和情感词的加入导致词共现率下降。

有情感联合和无情感联合这2种主题样例对比如表2所示。本文在主题数目设置为5的条件下以中文形式分别列举这2种方式主题的前8个词。

表2 IMDB数据集主题样例

Table 2 Topic samples of IMDB dataset

主题序号	有情感联合的主题	无情感联合的主题
1	浪费 最糟糕 弥补 可怕的 差的 业余的 色情 糟糕的	僵尸 血腥 僵尸(s) 砍刀 浪费 色情 赎回 废话
2	掠夺 谋杀 黑色 谋杀 伙伴 查尔斯 发现	新颖的 改编 精美 简 令人惊叹 杰出 出色
3	自由 纪录片 战争 社会 历史 政治 法国 深刻	决定 发现 杀人 同时 要求 找到 丈夫 谋杀
4	季节 情节(s) 情节 节目 迪士尼 喜爱 最喜爱 有趣	季节 节目 孩子 喜爱 笑话 爸爸 父母
5	爸爸 妈妈 杀死 僵尸 糟透了 愚蠢 死	情节 季节 系列 情节(s) 飞行员 行星 比赛 科学

从表2可以看出,有情感联合的主题表示样例大致可以将电影语料的主题概括为色情、犯罪、纪录、动画、恐怖5种类型,而无情感联合则稍显杂烩,较难概况其主题类型。该过程说明有情感联合可以学习更稀疏、更有意义的表示,其表达的主题关联强,其他主题关联弱的关键词较少,具有较优的表达主题语义的能力,主题解释性更强。相比无情感联合的主题表示,有情感联合的主题表示包含更多的情感词,有利于主题情感特征的获取,具有重要的实

际意义。由于引入背景术语先验知识,这2种主题表示样例中均减少了大量的“movie”、“film”等高频无显著主题表达意义的词的出现概率。

本文将构建SJST-VAE模型的文本语料的情感预测和主题的情感分布,利用生成的文档主题表征进行情感分类,通过单个主题的表征进行情感分布预测。由于本文主要侧重于挖掘主题的性能,因此不对情感分类准确率与其他模型进行对比。本文分别对10~100个主题数目进行模型训练,获得在不同

主题数目设置条件下的情感分类准确率,并累计计算5次情感预测准确率总和并取平均值,SJST-VAE模型情感预测准确率如图6所示。

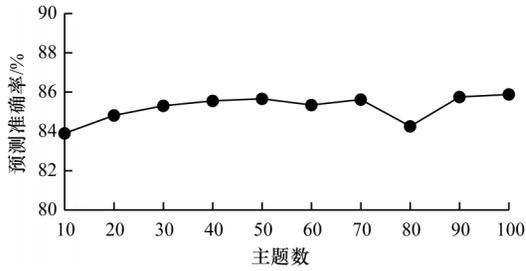


图6 SJST-VAE模型情感预测准确率

Fig.6 Sentiment prediction accuracy of SJST-VAE model

SJST-VAE模型在低维度文档表达条件下,仍具有较高的情感分类准确率,情感分类准确率并未随主题数增加而大幅波动,具有较强的稳定性。由于电影评论涉及较多情节内容,且其中包含的大量情感词不具有明确的实际褒贬意义,因此本文将在第6节的旅游具体应用中重点分析主题的情感分布过程,以及其如何用于指导主题的情感特征。

### 6 基于SJST-VAE模型的酒店用户画像构建

本文选择一组酒店评论文本集作为分析数据集<sup>[21]</sup>,为验证SJST-VAE模型在旅游推荐或游客群体画像中的实用性。该数据集中所有评论均来自于TripAdvisor.com的英国用户评价且每个评论文本均标注了情感极性,并区分了不同酒店级别和男女性别。据调查显示<sup>[22]</sup>,在较高星级酒店的选择上,男性和女性群体分别表现出不同的情感偏好和特征。通过挖掘分析获取不同群体的需求或喜好特点,进而推荐符合不同群体需求的酒店,成为提升游客体验和酒店运营的一个有效手段。

本文选取三星和四星这2种不同类型的酒店评论各6400条,每种类型酒店均包含男女性评论各3200条,并以此作为分析对象。整个数据集被划分为4个不同的特征数据集,如图7所示。同时,本文

将各数据集的80%作为训练集和20%作为测试集(正负评论数量均衡)。在训练过程中,本文设置词典大小为1000,批量大小设为50。主题个数设为10,既符合旅游酒店属性先验知识,也便于更细粒度了解用户需求和情感。

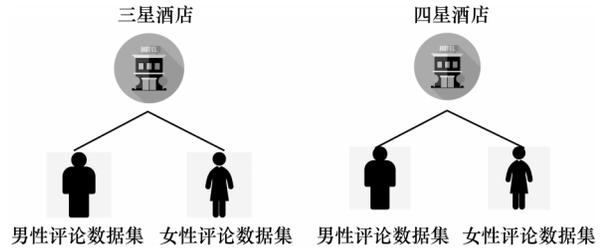


图7 不同特征数据集的划分

Fig.7 Division of different feature datasets

本文针对4种不同属性的酒店评论数据集分别进行主题情感的联合分析。SJST-VAE模型在4种不同属性的酒店评论数据集中情感预测准确率对比如图8所示。SJST-VAE模型在训练集和测试集的情感预测准确率均在90%以上,具有较高的准确率,验证了SJST-VAE模型在挖掘酒店用户评论特征进而获取情感预测的可行性。

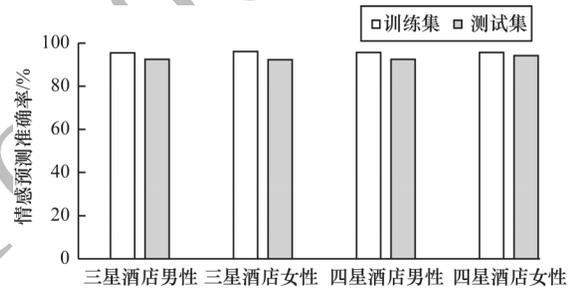


图8 在不同数据集上SJST-VAE模型的情感预测准确率对比

Fig.8 Sentiment prediction accuracy comparison of SJST-VAE model on different datasets

SJST-VAE模型分别对英国三星酒店男性和女性评论提取特征对比如表3、表4所示。

表3 三星酒店男性评论特征

Table 3 Feature of male reviews in Samsung hotels

%

序号	主题相关词	情感分布比例	
		负面	正面
1	敲打 走廊 味道 照片 床垫 睡眠 适应 按键	93.37	6.63
2	肮脏的 墙纸 发霉的 地毯 剥落 脏污的 瓷砖 脏的	99.72	0.28
3	地铁站 餐馆 城市 商店 中心 距离 公交车	1.11	98.89
4	非常精彩的 谢谢(s) 谢谢 迷人的 麻烦 高尔夫 享受的	0.62	99.38
5	休息室 套房 很好 大小 可口的高尔夫 美味的 喜欢	7.74	92.26
6	警报 电话 诱骗 高级 拒绝 退款 消防 联络	94.78	5.22
7	服务员 询问 果汁 空的 等待 土司 女服务员 牛奶	97.94	2.06
8	鱼薯条 鸡肉 午餐 品尝 开胃菜 干燥的 订购	96.16	3.84
9	客户 羞耻的 粗鲁的 电子邮件 等待 回应 经理 发送	97.99	2.01
10	赠送 美味的 餐桌 团队 午餐 生日 课程 课	10.19	89.81

表4 三星酒店女性评论特征

Table 4 Feature of female reviews in Samsung hotels

%

序号	主题相关词	情感分布比例	
		负面	正面
1	地铁 车站 距离 商店 伦敦 步行 便捷的一尘不染	0.83	99.17
2	退款 醒来 按键 联系 告知 通话 按键(s)	98.07	1.93
3	汽车 停车 停车场 出租车 购物 城市 火车 中心	76.44	23.56
4	发霉的 脏污的 墙纸 剥落 肮脏的 窗帘 地毯 家具	99.87	0.13
5	奇妙的 非常 感谢(s) 令人惊讶 建议 谢谢 麻烦 团队	0.19	99.81
6	高级的 旅馆 旅馆(s) 早餐 价值 舒适 的高效 附着	13.67	86.33
7	粗鲁 卡 建议 道歉 预定 令人震惊 态度 打电话	99.57	0.43
8	水疗 游泳池 按摩 浴缸 累 有用的 花园 娱乐 美丽	75.13	24.87
9	炒 豆沙 鸡蛋 沙发 鸡蛋(s) 不舒服的 冷冻 欺骗	95.98	4.02
10	鸡肉 薯条 盘子 冰淇淋 肉 点餐 午饭 课程	94.22	5.78

男性用户和女性用户均在房间噪音、内饰环境、餐饮、服务质量(包括入住办理、客房服务)上表现出负面倾向,如在内饰环境上的情感特征有“肮脏的”“发霉的”等,在服务质量上的情感特征有“羞耻的”“粗鲁的”“令人震惊的”等。在交通区位、休闲娱乐上,男性和女性均表现出一定的正面倾向,如交通区位上的情感特征有“便捷的”,休闲娱乐上的情感特征有“精彩的”“享受的”“谢谢”等。

女性的负面主题(7个)多于男性的负面主题(6个),可以推断女性在三星酒店消费中可能比男

性更为苛刻。相较于男性,女性更加注重细节,如房间内饰环境关键词上,女性增加了窗帘、家具的关注,在交通区位关键词上,女性增加了步行、停车、出租车、购物的关注,而男性则只是多了餐馆的关注。另外,在酒店休闲娱乐选择上,男女性也表现出不同的特点,如男性的休闲娱乐相关词有高尔夫等,女性的休闲娱乐相关词有水疗、游泳池、花园等。

SJST-VAE模型对英国四星酒店男性用户和女性用户评论提取的特征如表5、表6所示。

表5 四星酒店男性评论特征

Table 5 Feature of male reviews in four stars hotels

%

序号	主题相关词	情感分布比例	
		负面	正面
1	土司 队列 牛奶 移动 桌子 女服务员 注意 报价	97.87	2.13
2	地铁 车站 购物 距离 火车 城市 伦敦 酒吧	2.26	97.74
3	海滩 美味的 享受的 天气 宜人的 桑拿 游泳池 海	0.84	99.16
4	麻烦 谢谢 精彩 场地 令人惊讶 特别 高度 谢谢(s)	0.96	99.04
5	天花板 窗帘 衣柜 地毯(s) 厕所 地毯 墙 磨损的	98.54	1.46
6	电子邮件 详细的 解释 信用卡 回复 电话	98.66	1.34
7	肮脏的 高级 客店 打扰 孩子 提到 评论 游泳	96.09	3.91
8	服务 封闭的 眼镜 慢 差 可怕的高估 羞耻的	96.50	3.50
9	娱乐 享受的 态度 在里面 总是 团队 走 现在	75.79	24.21
10	鸡肉 开胃菜 午餐 牛排 女服务员 菜单 厨师 三明治	94.89	5.11

表6 四星酒店女性评论特征

Table 6 Feature of female reviews in four stars hotels

%

序号	主题相关词	情感分布比例	
		负面	正面
1	发生 电子邮件 解释 电话 讲 退款 经理 响铃	99.95	0.05
2	烤饼 蛋糕 三明治 奶油 蛋糕 下午 盘子 茶	88.91	11.09
3	染色 地毯 地毯(s) 床垫 墙 潮湿的 破碎的 脏的	99.90	0.10
4	距离 商店 购物 步行 配备 便捷的 地铁 车站	0.21	99.79
5	麻烦 谢谢 会场 团队 谢谢(s) 精彩的 婚礼 非常	1.22	98.78
6	护理 蒸汽 按摩 水疗 更换的 理疗 按摩 浴缸 桑拿	87.95	12.05
7	差 似乎 顾客 会议 希尔顿 娱乐 破旧的 圣诞节	96.18	3.82
8	汽车 飞行 停车场 公园 机场 睡眠 行李 噪音	55.31	44.66
9	超级 华丽 美丽 喜爱 海 花园 令人惊叹的 喜欢	0.86	99.32
10	灯 洗护用品 冰箱 中心 内陆 车站 街道 现代的	9.01	90.99

与三星酒店类似,男女性用户同样在房间噪音、内饰环境、餐饮、服务质量上表现出负面倾向,如在房间噪音上的情感特征有“打扰”“噪音”等,在内饰环境上的情感特征有“磨损的”“潮湿的”“破碎的”“脏的”等,在服务质量上的情感特征有“封闭的”“慢”“差”等。在交通区位、休闲娱乐、配套服务(如婚礼)上,男性女性均表现出一定的正面倾向,如交通区位上的情感特征有“便捷的”,休闲娱乐上的情感特征有“享受的”“宜人的”“精彩的”“喜欢”等。

男性关注的负面主题(7个)多于女性的负面(4个),可以推测四星酒店男性用户较三星酒店男性用户要求有所提高。在餐饮关键词上,男性多关注鸡肉、牛排,而女性更偏向于甜点如蛋糕、奶油和茶等。在房间内饰环境关键词上,男女性用户都关注了地毯、墙,男性相较于女性多了天花板、窗帘、衣柜、厕所的关注,女性则多了床垫的关注。在交通区位关键词上,男女性都关注了购物,男性较女性多了酒吧的关注,女性较男性则多了步行的关注。在休闲娱乐上,男性用户评论的关键词有海滩、海、美味、桑拿、游泳池等,女性用户评论的关键词有护理、蒸汽、水疗、按摩、海、花园等。

通过以上分析,酒店运营者可以从男性和女性用户在不同星级酒店消费过程中所关注的内容和相应感受,获取男性女性用户的不同特征,进而有针对性地为客户偏好层次上进行酒店或房间的推荐。通过对男性女性用户所表现出的负面主题和情感特点进行分析,以促使酒店管理者发现内部不足进而提出改进措施。相对主题和情感的割裂分析,针对主题情感分布的挖掘更具有实际应用价值。

SJST-VAE模型是基于正负分布均衡的酒店评论数据集,但是表3~表6所呈现出的用户负面主题却明显大于正面主题,这或许是由于用户的表达习惯所决定的。在评价事物时,负面信息的可诊断性要强于正面信息,消费者会赋予负面信息更高的权重或注意力。对于用户是否习惯于在负面主题的表达更加具象,而在正面主题的表达更加笼统如“太美了、太舒服了、非常享受等”,从而导致模型挖掘到的负面主题方面多于正面,还需要后续大量的实验进行佐证。

## 7 结束语

为充分捕捉用户细粒度的意见,本文构建基于变分自编码的神经网络训练模型SJST-VAE。利用先验知识和情感标签辅助主题的训练和生成,基于截断高斯模型,构造更适合Dirichlet过程的神经变分推断形式,其中截断高斯模型中的截断结构能够有效地捕获离散数据中的相关性,适用于主题分类数据的分析。实验结果证明,SJST-VAE模型能够利用主题分布实现情感分类的预测。酒店运营者通过SJST-VAE模型获取用户群体的情感偏好或舆情报告,有助于制定详实可靠的改进措施。下一步将把本文模型应用在旅游领域的精准推荐系统中,以实现在不同应用场景下信息的融合与扩展。

## 参考文献

- [ 1 ] 何力,郑灶贤,项凤涛. 基于深度学习的文本分类技术研究进展[J]. 计算机工程,2021,47(2):1-11.  
HE L, ZHENG Z X, XIANG F T, et al. Research progress of text classification technology based on deep learning[J]. Computer Engineering, 2021, 47(2): 1-11. (in Chinese)
- [ 2 ] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.
- [ 3 ] LIN C, HE Y. Joint sentiment/topic model for sentiment analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, USA: ACM Press, 2009: 375-384.
- [ 4 ] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2011: 815-824.
- [ 5 ] ALAM M H, RYU W-J, LEE S. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews[J]. Information Sciences, 2016, 339: 206-223.
- [ 6 ] 郝洁,谢珺,苏婧琼,等. 基于词加权LDA算法的无监督情感分类[J]. 智能系统学报,2016,11(4):539-545.  
HAO J, XIE J, SU J Q, et al. An unsupervised approach for sentiment classification based on weighted latent dirichlet allocation[J]. CAAI Transactions on Intelligent Systems, 2016, 11(4): 539-545. (in Chinese)
- [ 7 ] NGUYEN T H, SHIRAI K. Topic modeling based sentiment analysis on social media for stock market prediction[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China: [s. n.], 2015: 1354-1364.
- [ 8 ] LIANG Q, RANGANATHAN S, WANG K, et al. JST-Rr model: joint modeling of ratings and reviews in sentiment-topic prediction[EB/OL]. [2021-03-25]. <https://arxiv.org/abs/2102.11048v1>.
- [ 9 ] 袁非牛,章琳,史劲亭,等. 自编码神经网络理论及应用综述[J]. 计算机学报,2019,42(1):203-230.  
YUAN F N, ZHANG L, SHI J T. Theories and applications of auto-encoder neural networks: a literature survey[J]. Chinese Journal of Computers, 2019, 42(1): 203-230. (in Chinese)
- [ 10 ] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Supervised topic models[EB/OL]. [2021-03-25]. <https://arxiv.org/pdf/1601.00670.pdf>.
- [ 11 ] BOWMAN S R, VILNIS L, VINYALS O, et al. Generating sentences from a continuous space[EB/OL]. [2021-03-25]. <http://de.arxiv.org/pdf/1511.06349>.
- [ 12 ] XU W, SUN H, DENG C, et al. Variational autoencoder for semi-supervised text classification[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2017: 1-10.
- [ 13 ] SRIVASTAVA A, SUTTON C. Autoencoding variational inference for topic models[EB/OL]. [2021-03-25]. <https://arxiv.org/abs/1703.01488v1>.
- [ 14 ] TOMCZAK J, WELLING M. VAE with a VampPrior[EB/OL]. [2021-03-25]. <https://arxiv.org/pdf/1705.07120.pdf>.

(下转第294页)

(上接第 287 页)

- [15] TAKAHASHI H, IWATA T, YAMANAKA Y, et al. Variational autoencoder with implicit optimal priors[C]// Proceedings of AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2019; 5066-5073.
- [16] 黄佳佳, 李鹏伟, 彭敏, 等. 基于深度学习的主题模型研究[J]. 计算机学报, 2020, 43(5): 75-103.  
HUANG J J, LI P W, PENG M, et al. Review of deep learning-based topic model [J]. Chinese Journal of Computers, 2020, 43(5): 75-103. (in Chinese)
- [17] HOANG T, LE H, QUAN T. Towards autoencoding variational inference for aspect-based opinion summary[J]. Applied Artificial Intelligence, 2019, 33(9): 796-816.
- [18] EISENSTEIN J, AHMED A, XING E P. Sparse additive generative models of text[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. Washington D. C. , USA: IEEE Press, 2011; 1041-1048.
- [19] MIAO Y, YU L, BLUNSOM P. Neural variational inference for text processing [C]//Proceedings of International Conference on Machine Learning. Washington D. C. , USA: IEEE Press, 2016; 1727-1736.
- [20] TripAdvisor UK reviews gendered data sets with equal numbers of all five ratings[EB/OL]. [2021-03-25]. [https://figshare.com/articles/dataset/TripAdvisor\\_reviews\\_of\\_hotels\\_and\\_restaurants\\_by\\_gender/6255284](https://figshare.com/articles/dataset/TripAdvisor_reviews_of_hotels_and_restaurants_by_gender/6255284).
- [21] LI Q, LI S, ZHANG S, et al. A review of text corpus-based tourism big data mining[J]. Applied Sciences, 2019, 9(16): 3300.
- [22] JUWAHEER T D. Gender bias in hotel guests' perceptions of service quality: an empirical investigation of hotels in Mauritius[J]. E-review of Tourism Research, 2011, 9(5): 1-10.