计算机科学
COMPUTER SCIENCE

突发事件中网络评论的情感-主题随时间的演变研究

史伟, 付月

引用本文

史伟, 付月. 突发事件中网络评论的情感-主题随时间的演变研究[J]. 计算机科学, 2022, 49(11A): 211000193-6.

SHI Wei, FU Yue. Study on Evolution of Sentiment-Topic of Internet Reviews with Time in Emergencies [J]. Computer Science, 2022, 49(11A): 211000193-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000193

突发事件中网络评论的情感-主题随时间的演变研究

史 伟[1] 付 月[2]
1 湖州师范学院经济管理学院 浙江 湖州 313000
2 湖州学院经济管理学院 浙江 湖州 313000

摘 要 网络评论的情感主题演变分析对突发事件中网络舆情的控制极具价值。针对情感主题动态性的特点,构建一个基于LDA的情感主题模型,通过对时间与主题和情感的联合建模来分析情感主题随时间的演变,推导了基于Gibbs抽样过程的推理算法,最后通过微博突发事件数据集的分析结果显示了联合模型较高的准确性和情感主题随时间演变过程中良好的应用性。

关键词:时间感知情感主题模型;时间序列;趋势分析;情感分析
中图法分类号 TP391.1

Study on Evolution of Sentiment-Topic of Internet Reviews with Time in Emergencies

SHI Wei[1] and FU Yue[2]
1 School of Economics and Management, Huzhou Normal University, Huzhou, Zhejiang 313000, China
2 School of Economics and Management, Huzhou University, Huzhou, Zhejiang 313000, China

Abstract The analysis of sentiment topic evolution is of great value to the control of network public opinion in emergencies. According to the dynamic characteristics of sentiment topics, this paper constructs a sentiment topic model based on LDA, analyzes the evolution of sentiment topics with time through the joint modeling of time, topic and sentiment, deduces the reasoning algorithm based on Gibbs sampling process, and finally puts forward the analysis results of product reviews and microblog emergency data sets, which shows that the joint model has good accuracy and good applicability in the process of time evolution.

Keywords Time-aware sentiment-topic model(TST), Time series, Trend analysis, Sentiment analysis

1 引言

主题建模和情感分析是处理文本数据的两个常用任务。前者处理主题的提取(它是关于什么的?),后者是关于情绪和意见分类(基本观点是什么?)。这两个任务是互补的,在某种程度上,情感通常是关于主题的,而主题往往是主观立场的基础。这就是为什么主题和情感应该被联合提取和分析。近年来,联合主题情感建模作为一项独立的文本挖掘任务应运而生。文献[1-5]已经做了一些有用的工作,但这些工作大多是静态地提取主题的情感,而忽略了文本数据的动态性质。还有些工作如,文献[6-8],只专注于分析主题层面的内容演变,而忽略了主题情感的相关性。基于这一观察结果,我们提出了一种基于主题模型的主题情感关联性提取方法,以获得主题情感相关性及随时间的演变过程。

该模型产生了3个层次的输出:主题、主题情感和主题情感随时间的演化。它首先作为一个传统的主题发现模型,能够从文档集合中提取隐藏的主题结构。其次,对主题和情感(对每个提取主题的总体情感)之间的关联进行建模。最后,提供了一个有效的工具来跟踪和可视化主题情感关联的强度。所有这些信息都是同时提取的,不需要任何后期处理。

所提方法有3个主要特点,这3个特点是其他文献所不能共同解决的。首先,时间与主题和情感共同建模,这使得能够捕捉主题情感随时间的演变。其次,针对的是整个数据而不是单个文档,一次提取主题特定情感,从而可以提供主题情感相关性的整体视角。最后,在不同的情感极性下,不需要进行后处理来匹配相似的主题。

将文中模型与其他先进的主题情感模型JST和ASUM进行了比较。在两个不同的数据集(包括产品评论和新闻文章)上的实验,证明了我们的模型在提取准确的主题情感时间关联性方面的有效性。本文第2节概述了相关工作;第3节介绍了研究思路和方法;第4节和第5节给出了实验、结果和讨论;最后总结全文。

2 研究现状

2.1 联合主题情感建模

主题和情感建模任务与通常针对某个主题表现出的情感程度相关。为了对主题情感连接进行建模,很多研究使用主题模型:基于词共现模式从文本中发现低维结构(主题)的统计模型。早期的主题模型如LDA[9]和PLSA[10],主要集中于提取同质的主题,但最近这些模型被扩展到捕捉文本的其他方面,比如情感。以联合情感主题模型(JST)[2,11]为例,他们已经构建了不同情感标签下的主题抽取方法,通过在主题层之前插入一个新的情感层来扩展LDA。因此要为文档生成一个词语,首先绘制情感标签 s,然后根据 s 绘制主题。Xu

基金项目:国家社会科学基金一般项目(20BXW013)
This work was supported by the General program of National Social Science Foundation of China(20BXW013).
通信作者:史伟(shiwei@zjhu.edu.cn)

等[12]在 JST 的基础上提出了考虑用户特征的主题情感联合(JUST)模型,该方法将用户特征加入模型,以文档所对应的用户特征的线性函数作为文档-情感分布的先验,由此得到具有不同特征的用户群体的情感倾向。反向 JST[2]是 JST 的一个变体,其中情感层和主题层的顺序是颠倒的。其他不同的模型如主题-情感混合(TSM)[3]、情感 LDA[1]、情感统一模型(ASUM)[4]和具有分解先验的情感主题模型(STDP)[5],除基于 PLSA 的 TSM 外,所有这些模型都基于 LDA。

2.2 主题随时间的演变

文档通常是随着时间的推移而收集的(在线讨论、新闻、电子邮件等),因此它们的内容可能也会随着时间的推移而变化。这里我们关注定量演变,即在某个时间戳 t 讨论某个主题的数据量。TOT[8]是一个基于 LDA 的定量主题演化模型,通过每周在 Twitter 上列出关于 COVID-19 讨论最多的话题,并每周监控话题的演变。在文献[7]中,LDA 模型被用来通过计算每个时间戳上与每个主题相关联的文档数量来捕捉主题随时间的变化。也有学者提出了动态主题模型(DTM)[6]来模拟主题词分布随时间的变化,将 Dirichlet 多项式或 Pitman-YR 过程等概率混合模型与 Gibbs 采样和随机变分推理等近似推理方法相结合,分别考虑了 DTM 的动态性和可扩展性。国内学者 Jiang 等[13]构建了基于 LDA 的产品在线评论主题演化分析模型,从主题标签、主题热度和主题词热度 3 个层面挖掘海量在线评论的主题演化,发现了一系列新规律,取得了不错的效果。

2.3 联合主题情感随时间的演变

主题情感演变的建模是一个相对较新的问题,在文献中较少涉及。Mei 等[3]是最早处理这个问题的团队之一,利用先前提出的 TSM 模型提取主题情感关联,定量演化的特征是在同一时间戳中一个主题和一个情感标签指定的词语数。文献[14]提出了一个基于流形学习的模型来研究在线新闻领域中话题情感关联及其随时间的演化,该模型能在低维空间中直观地反映主题的情感动态。文献[3]和文献[12]中的方法最接近本文研究,然而我们的模型不同于文献[3],因为它不需要经过后处理来推断时间演化。文献[12]中的模型是基于 PLSA 的,众所周知 PLSA 有许多缺点,如过度拟合学习数据和由于大量学习参数而导致的高推理复杂度[9]。文献[15]提出了一种基于动态主题情感演化模型的舆情信息分析方法,将改进的 TF-ID 和 K-Means 聚类方法相结合提取主题词,形成主题-情感匹配表,引入时间节点,利用 PMI 和情感词典进行动态情感演化分析,具有一定实效性。He 等[16-17]在 JST 模型的基础上引入了动态 JST,以捕捉随时间推移的定性主题演变,他们的方法类似于 DTM[6],其中每个时间戳的模型都是从上一个时间戳的模型派生出来的。

本文方法不同于许多基于模型适应性(定性演变)的主题模型,我们基于著名的 LDA 模型,既具有相同的优点,特别是明确使用 Dirichlet 超参数来平滑多项式分布,这些超参数可以指导联合主题情感的发现,优化情感主题模型[2,4-5],又通过对时间与主题和情感的联合建模来分析情感主题随时间的演变,对突发事件中网络舆情的分析具有一定价值。

3 研究思路与方法

3.1 模型图和符号解释

本节中将描述时间感知情感主题(Time-aware Sentiment-Topic, TST)模型。本文方法是对主题情感关联性以及它们随时间定量演化过程进行建模,是在以下一些模型的基础上提出的:1)时间没有与主题和情感一起建模[1-5,13-14];2)对每个文档的主题特定情感分别进行评估[1,2,4-5];3)来自不同情感极性的相似主题不会自动匹配[1-5]。

为了解决这些问题,基于 3 个主要特征提出了一个新颖的主题模型:1)时间与主题和情感共同建模,定量分析了主题情感随时间的演变;2)针对整个数据而不是单个文档提取主题特定情感,从而提供主题情感相关性的总体视图;3)不需要后处理来匹配不同情感极性下的主题,因为同一主题在词语上有多个分布,每个情感极性对应一个分布。

通过添加两个新的层 s 和 t 来扩展 LDA 模型,以便分别捕捉情感和时间(见图 1)。我们的方法是建立在传统的主题建模假设之上:学习集合中的每个文档都是主题的混合(主题的多项式分布)。此外假设每个主题都有多个方面,每个情感极性对应一个,因此在词语上有多个多项式分布。最后假设主题-情感关联的"强度"会随着时间的推移而演变。

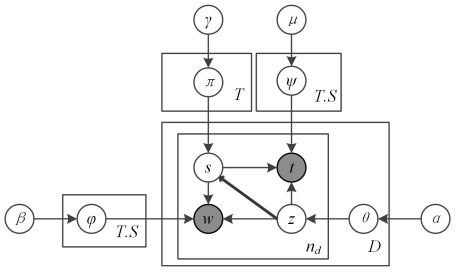

图 1 时间感知情感主题(TST)图形模型

Fig. 1 Graphical model of time-aware sentiment-topic model (TST)

学习数据中的文档必须标注时间(如创建日期)。首先对时间进行离散化,每个文档都会收到一个离散的时间戳标签(如年、月、日)。在学习步骤中,使用变量 t 捕捉时间模态,然后使用时间戳 ψ 上的多项式分布捕捉主题情感演化。本文其余部分使用的符号如表 1 所列。

表 1 符号解释

Table 1 Symbol interpretation

符号	解释
D	文档数量
V	词汇量
T	主题数量
S	情感标签数量
H	时间戳数量
θ	$[\theta_d]$:$D\times T$ 主题文档特定分布矩阵
φ	$[\varphi_{z,s}]$:$T\times S\times V$ 主题情感特定词分布矩阵
Π	$[\pi_z]$:$T\times S$ 主题特定情感分布矩阵
ψ	$[\psi_{z,s}]$:$T\times S\times H$ 特定于主题情感对的时间分布矩阵
n_d	文档 d 中的词语数
$n_{d,j}$	文档 d 中受主题 j 影响的词语数
n_j	主题 j 影响的词语数
$n_{j,k}$	影响话题 j 和情感 k 的词语数
$n_{i,j,k}$	词语 i 受主题 j 和情感 k 影响的次数
$n_{j,k,h}$	时间戳为 h 的词语受主题 j 和情感 k 影响的次数
n^{-p}	在当前文档的 p 位置排除词语的数量变量

3.2 生成过程

TST 是一个词语、情感和时间戳完全生成的模型。其生成过程如下:

(1)绘制 $T\times S$ 多项式 $\varphi_{z,s}\sim \mathrm{Dir}(\beta)$。

(2)绘制 $T\times S$ 多项式 $\psi_{z,s}\sim \mathrm{Dir}(\mu)$。

(3) 绘制 T 多项式 $\pi_z \sim \mathrm{Dir}(\gamma)$。

(4)对于每一个文档 d，绘制一个多项式 $\theta_d \sim \mathrm{Dir}(\alpha)$，然后对于文档 d 中的每个词语 w_i：

1)绘制一个主题 $z_i \sim \theta_d$；

2)绘制一个情感标签 $s_i \sim \pi_{z_i}$；

3)绘制一个词语 $w_i \sim \varphi_{z_i,s_i}$；

4)绘制一个时间戳 $t_i \sim \psi_{z_i,s_i}$。

通过研究 TST 的图形模型和生成过程，可以注意到同一文档中的不同词语可能会生成不同的时间戳，但是文档中的所有词语都应该有相同的时间戳。实际上这并不是一个真正的问题，因为 TST 在情感主题动态建模方面仍然是有效的。然而由于时间模态涉及到主题发现，可能会影响主题的同质性，因为时间模态被假定为与词语模态具有相同的“权重”。为了解决这个问题，我们采用了与 TOT 模型[8]和组主题模型[18]相同的策略，引入了一个平衡超参数来平衡词语和时间在主题发现中的贡献。一个自然的设置是使用词语数 n_d 的倒数作为平衡超参数，在计算后验分布时考虑这个超参数。

3.3　推理过程

吉布斯抽样(Gibbs)是主题模型[8]中常用的参数估计(推理)方法，我们采用这种方法是因为它通常产生相对简单的算法。由于篇幅的限制，这里只给出最终公式。文献[19]提供了通过吉布斯抽样对 LDA 进行推断的详细推导，对 TST 的派生是以相同的方式执行的。

(1)联合分布。使用 Bayes 条件独立性规则，词语、主题、情感和时间戳的联合概率可以计算如下：

$$p(w,t,s,z|\alpha,\beta,\gamma,\mu)=p(w|s,z,\beta)\cdot p(t|s,z,\mu)\cdot p(s|z,\gamma)\cdot p(z|\alpha) \tag{1}$$

式(1)中的第一项是通过对 φ 进行积分得到的。

$$p(w|s,z,\beta)=\left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{T\cdot S}\prod_j\prod_k\frac{\prod_i\Gamma(n_{i,j,k}+\beta)}{\Gamma(n_{j,k}+V\beta)} \tag{2}$$

其中，Γ 表示伽马函数，下标 i,j,k,h 分别用于循环词语、主题、情感和时间戳。

式(1)的第二项是通过在 ψ 上积分得到的。

$$p(t|s,z,\mu)=\left(\frac{\Gamma(H\mu)}{\Gamma(\mu)^H}\right)^{T\cdot S}\prod_j\prod_k\frac{\prod_h\Gamma(n_{j,k,h}+\mu)}{\Gamma(n_{j,k}+H\mu)} \tag{3}$$

式(1)中的其余各项分别通过在 π 和 θ 上积分得到。

$$p(s|z,\gamma)=\left(\frac{\Gamma(\sum_k\gamma_k)}{\prod_k\Gamma(\gamma_k)}\right)^T\prod_j\frac{\prod_k\Gamma(n_{j,k}+\gamma_k)}{\Gamma(n_j+\sum_k\gamma_k)} \tag{4}$$

$$p(z|\alpha)=\left(\frac{\Gamma(\sum_j\alpha_j)}{\prod_j\Gamma(\alpha_j)}\right)^T\prod_d\frac{\prod_j\Gamma(n_{d,j}+\alpha_j)}{\Gamma(n_d+\sum_j\alpha_j)} \tag{5}$$

(2)后验分布。后验分布估计是通过抽样变量 z,s 给定所有其他变量。我们使用上标 $-p$ 表示不包括当前文件 d 的位置 p 处的词语数量。后验概率可由联合概率得出，如下所示：

$$p(s_p=k,z_p=j|w,t,s^{-p},z^{-p},\alpha,\beta,\gamma,\mu)\propto\frac{n_{d,j}^{-p}+\alpha_j}{n_d^{-p}+\sum_j\alpha_j}\cdot\frac{n_{w_p,j,k}^{-p}+\beta}{n_{j,k}^{-p}+V\beta}\cdot\frac{n_{j,k}^{-p}+\gamma_k}{n_j^{-p}+\sum_k\gamma_k}\cdot\frac{n_{j,k,t_p}^{-p}+\mu}{n_{j,k}^{-p}+H\mu} \tag{6}$$

平衡超参数 $\frac{1}{n_d}$ 作为式(6)最后一项的指数幂引入，然后使用从马尔可夫链获得的样本来估计分布 φ,θ,π 和 ψ，如下所示：

$$\begin{cases}\varphi_{j,k,i}=\dfrac{n_{i,j,k}+\beta}{n_{j,k}+V\beta},\theta_{d,j}=\dfrac{n_{d,j}+\alpha_j}{n_d+\sum_j\alpha_j}\\ \pi_{j,k}=\dfrac{n_{j,k}+\gamma_k}{n_j+\sum_k\gamma_k},\psi_{j,k,h}=\dfrac{n_{j,k,h}+\mu}{n_{j,k}+H_\mu}\end{cases} \tag{7}$$

TST 推理的步骤如算法 1 所示。

算法 1　TST 推理

Require: α,β,γ,μ, T

1. Initialize matrices Φ,Θ,Π,Ψ.
2. for iteration c=1 to nbGibbsIterations do
3. for document d=1 to D do
4. for p=1 to n_d do
5. Exclude word w_p from d and update count variables
6. Sample a topic and a sentiment label for word w_p using Equ. 6
7. Update count variables with new topic and sentiment label
8. end for
9. end for
10. end for
11. Update matrices Φ,Θ,Π,Ψ with posterior estimates using Equ. 7

此推理算法的输入数据分别是词语、主题、情感和时间戳的超参数(α,β,γ,μ)，输出是更新矩阵 Φ,Θ,Π,Ψ。

3.4　融合情感本体库

以情感词典的形式来指导情感发现。当对一个词语的情感进行抽样时，会引入情感词典(算法 1 的第 6 行)。如果一个词语出现在词典中，它就会受到词典中相应的情感标签的影响，否则使用式(6)生成情感标签。文献[1-2,5]中也采用了这种策略。这里使用已经构建的模糊情感本体库来评估我们的方法，在前期研究中已详细论述情感本体的构建过程[20]，创建了可用于在线评论情感分析的情感词本体库。主要创新之处是将情感本体划分为评价词本体和情感词本体，利用模糊理论和知网(HowNet)模型构建情感本体的基本模型。根据评价词和情感词各自的特点，运用模糊化处理和语义相似度的相关理论，分别对评价词本体和情感词本体的情感类型和强度进行了相应处理。情感本体形式如下：

FEO=((18;开心;happy;adj;张三;知网 2007 版情感分析用词语集),(快乐;愉快),(高兴;1.00))

最终的情感本体收录 8 952 个词条，各类情感(2 种评价类和 8 种情感类)统计如表 2 所列。

表 2　各情感类词汇数量

Table 2　Number of sentiment words

情感类	G(好)类评价词	B(坏)类评价词	期待	高兴	喜爱	惊讶	焦虑	悲伤	生气	讨厌
词汇数	3715	3147	170	395	339	65	271	220	201	429

各情感类词汇分别赋予了相应的情感类和情感强度值，情感强度取值范围为[0,1]。情感有正面和负面之分，即情感极性。上述 10 类情感中，G 类评价词、期待、愉快、喜爱属于正面情感，而 B 类评价词、悲伤、生气和讨厌则属于负面情感，惊讶和焦虑在不同的语境下既可能表现为正面也可能为负面。

4　实验过程

4.1　评价框架

评价 TST 模型至少涉及两个方面：情感主题关联和情感主题随时间的演变。本文提出通过模型结果和实际数据的比较来评估这两方面。为此将采用这样的数据集：其中每个文档都用主题、情感和时间进行了标注。然后对于每一个主题情感对，通过合并标注为情感 s 的主题 z 的所有文档来计算

词语 $p(w|s,z)$ 上的“真实”(观察)分布。通过计算标注为情感 s 的主题 z 的文档数,来计算每个主题在情感 $p(s|z)$ 上的实际分布。最后,用同样的方法计算主题情感对随时间 $p(t|s,z)$ 的实际分布。基于真实数据定义两个独立的评价指标:主题情感关联准确度 Q_s 和主题情感演变准确度 Q_t。这些措施是基于“估计”和“真实”之间距离的计算,一般分为两个步骤:主题匹配和评价措施。

(1)主题匹配

为了简单起见,这里假设二元情感模式:正面情感和负面情感。设 r,e 为真实,分别估计主题。基于词汇表 φ_r 和 φ_e 上主题分布之间的 KL 散度的计算,每个主题 r 与主题 e 进行匹配。由于 KL 散度不是一个距离度量,所以这里使用 KL 距离(KLD)代替。对于两个多项式分布 P 和 Q,计算如下[21]:

$$KLD(P,Q)=KL(P\|Q)+KL(Q\|P)$$
$$=\sum_i\left[(P(i)-Q(i))\cdot\log\frac{P(i)}{Q(i)}\right] \tag{8}$$

匹配过程通过迭代选取 KLD 值最小的主题对来实现。一般过程分为两步:1)在正负极性下分别匹配真实主题和估计主题;2)如果在前一步中 e_p 和 e_n 与同一个真实主题匹配,则来自估计正面主题的每个主题 e_p 与来自估计负面主题的一个主题 e_n 匹配。这种双重匹配只对 JST 和 ASUM 模型是必需的,在 TTS 中它由模型自动提供。

(2)评价措施

设 M 是上一步的结果(M 包含匹配的主题对,而不考虑极性)。每对主题 $(r,e)\in M$ 以情感分布为特征。这种分布的计算是针对每个模型的,对于 TST 它直接由模型(分布 π)产生,对于 JST 和 ASUM,$p(s|z)$ 的获得方式与实际分布计算类似,但带有新的(估计的)标注。每个文档 d 都用情感和主题最大化概率 θ_d 重新标注。第一个评估指标 Q_s(主题情感关联准确度)是匹配主题对的真实分布和估计 π 分布之间的平均 KL 距离:

$$Q_s=\frac{1}{T}\cdot\sum_{(r,e)\in M}KLD(\pi_r,\pi_e) \tag{9}$$

第二个评价指标 Q_t 是基于估计的主题情感随时间分布(ψ)的计算,这个信息由 TST 直接生成。对于 JST 和 ASUM,我们使用与文档相关联的实际时间戳来估计 ψ 分布。最后,主题情感时间关联准确度 Q_t 是匹配主题对的真实分布和估计 ψ 分布之间的平均 KL 距离:

$$Q_t=\frac{1}{T}\cdot\sum_{(r,e)\in M}KLD(\psi_r,\psi_e) \tag{10}$$

4.2 实验数据

本文使用的数据集为“贵州公交车坠湖”微博数据集。“贵州公交车坠湖”是指 2020 年 7 月 7 日 12 时 12 分,贵州省安顺市一辆 2 路公交车从安顺火车站驶向客车东站,在途经虹山湖大坝中段时,冲破石护栏坠入湖中。公众对这一事件在微博中展开了热烈讨论,形成了巨大舆情。微博数据集是 2020-07-07—2020-07-14 期间通过“贵州安顺”“贵州公交车坠湖”等关键词收集微博上关于该事件的相关微博评论文本,已在前期研究中对该数据集进行过情感标注和分析。这个数据集按照如下步骤进行了规范和解析,预处理数据统计如表 3 所列。

(1)对在空白边界上的个别词进行分离;

(2)从词语中去除所有非文字的数字字符,例如逗号、破折号等;

(3)去除 1208 个标准停用词,包括常见的一些动词形式;

(4)为了避免垃圾信息和其他一些不相关的信息,从数据集中过滤掉额外的链接,如含有“http:”或者“www.”的表达和用户的名字(用符号@标志的);

(5)移除“回复”“转发微博”等词和转发的内容(只是转发没有增加任何评论的帖子)。

表 3 所用数据集的统计信息

Table 3 Statistics of data sets used

数据集	类型	D	V	标注	时间标记
“贵州公交车坠湖”微博数据集	新闻/微博评论	146518	456725	主题、情感、时间	日

基于 GibbsLDA++7 的代码实现了 TST 模型。考虑两种情感极性:正面和负面。对已构建的情感本体库中的情感类做了相应处理,将 G 类评价词、期待、愉快、喜爱和惊讶定为正面情感,将 B 类评价词、悲伤、生气、讨厌和焦虑定为负面情感。对于这两个数据集,将主题数 T 设置为 9,对称超参数 α,β 和 μ 分别设置为 $\frac{40}{T}$,0.04 和 0.01。实验表明,TST 对参数 μ 不敏感,即使 μ 值很低,时间稀疏性也不是什么大问题。为了评估,当 γ_{pos}. 为变量时,超参数 γ_{pos}. 设置为 1。实验表明主题情感模型 TST,JST 和 ASUM 对这些参数的取值不敏感,而对它们的比值 $\frac{\gamma_{\text{neg}}}{\gamma_{\text{pos}}}$ 敏感,这里用 γ_{ratio} 表示。所有结果均在 *Gibbs* 采样器第 400 次迭代时得到。

5 实验结果

5.1 主题情感提取

时间感知情感主题模型(TST)的首要任务就是提取主题情感关联。图 2 显示了使用主题情感准确性指标 Q_s(参见第 4 节)对 TST,JST 和 ASUM 模型的定量评价,值越小越好。这一结果是在微博评论数据集上通过改变 γ_{ratio} 得到的。在 Q_s 方面,ASUM 给出的结果最好($Q_s=0.4,\gamma_{\text{ratio}}=400$),其次是 TST($Q_s=1.86$),然后是 JST($Q_s=2.23$)。这项实验表明,当把句子作为连贯单位(如 ASUM)而不是词语(如 TST 和 JST)处理时,主题情感模型更有效。研究还发现,与 JST 相比,TTS 和 ASUM 对初始化步骤(将词语随机分配给主题和情感)的敏感度更低。

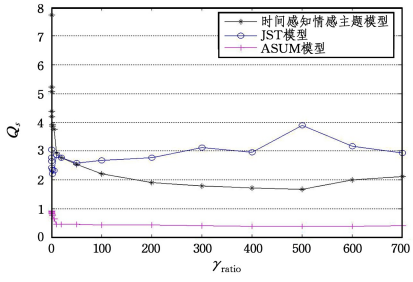

图 2 微博数据集上的主题情感关联准确性 Q_s

Fig. 2 Topic sentiment association accuracy(Q_s) on Weibo dataset

然而,TST 在提取主题和主题的情感方面仍然很有效。表 4 列出了使用 TST 模型从“贵州公交车坠湖”微博数据集中选取的主题示例,其中最后一行表示特定于主题的总体

情感概率($\pi_z(s)$)。这个数据集的 γ_{ratio} 比值设置为 300，主题由两种情感极性下最有可能出现的词语的有序列表来表示。从表中可以看出提取的主题显然是顽固的，在每一种情感标签下最有可能的词是相当连贯和情绪化的。例如主题 z_3(学生)在正极性下被正面描述("希望""幸运"等)，同样的主题在负极性下用否定的词来描述("惨""难受"等)。

表 4　"贵州公交车坠湖"(z_1 — z_4)数据集中所选主题的关键词

Table 4　Keywords of selected topics in the "Guizhou bus falling into the lake"(z_1 — z_4) dataset

z_1：公交车	z_2：司机	z_3：学生	z_4：调查结果
正面 负面	正面 负面	正面 负面	正面 负面
行驶 坠河	加速 恐怖	希望 揪心	通报 难受
很棒 侧翻	期盼 报复	抢救 惨	意外 伤亡
方便 强制	感谢 恐惧	同情 受伤	赔偿 诡异
智能 坠湖	好　故意	救援 害怕	真相 气愤
调度 撞坏	适合 蓄意	考试 溺水	报警 憎恨
享受 失控	平安 严惩	平安 无辜	期待 郁闷
视频 诡异	同情 危害	幸运 难受	勇敢 荒唐
0.45 0.55	0.36 0.64	0.73 0.27	0.21 0.79

5.2　主题情感随时间的演变

TST 的第二个目标是模拟主题情感随时间的演变。图 3 给出了 Q_t 测量值随 γ_{ratio} 比值的变化(Q_t 值越小越好)，Q_t 用来测量模型如何随时间获得准确的主题情感关联。图 3 显示 TST 模型在 Q_t 度量方面显著优于 JST 和 ASUM，TST 得到的最佳结果是 $Q_t=1.78$，而 JST 和 ASUM 的 Q_t 分别为 3.88 和 3.69。这个实验表明在 TST 的建模过程中加入时间信息，有助于更准确地提取主题-情感-时间关联。

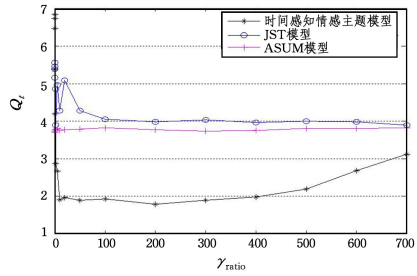

图 3　微博数据集上的主题-情感-时间关联准确性 Q_t

Fig. 3　Topic-sentiment-time association accuracy(Q_t) on Weibo dataset

仍以"贵州公交车坠湖"微博数据集为例对 TST 模型进行实证研究。为便于解释，运用同属于主题 z 和情感 s 的文档数量来度量主题 z 和情感标签 s 的主题-情感演变。在每个时间戳 t 的文档数记为 $nbDocs_{z,s}(t)$，计算如下：

$$nbDocs_{z,s}(t)=\phi_{z,s,t}\cdot\pi_{z,s}\cdot topicSize(z) \tag{11}$$

其中，$topicSize(z)$是使用最大概率分配给主题 z 的文档数。

图 4 给出了来自"贵州公交车坠湖"微博数据集的一组主题的估计演变。

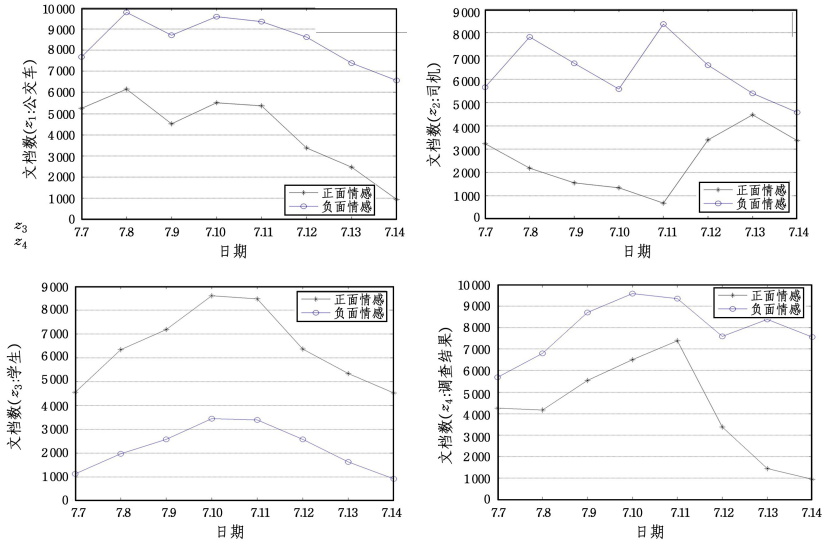

图 4　使用 TST 模型估计"贵州公交车坠湖"微博数据集上随时间的主题情感演变(文档数)

Fig. 4　Estimation of topic sentiment evolution(number of documents) over time on the Weibo dataset of "Guizhou bus falling into the lake" using TST model

深入研究图 4 中的主题情感演变，可以得出以下信息。

(1)主题 z_1 指的是关于"公交车"的相关新闻和微博评论。在事故发生后的 8 天时间里关于贵州公交车的总体情感比较负面，可以发现"失控""诡异""坠湖"等词语的使用(见表 4)；事故发生第二天公众对该公交车的不正常行驶轨迹表示震惊和质疑，讨论数达到高峰，随着事故的调查，该主题的讨论渐趋下降。

(2)主题 z_2 指的是关于"司机"的相关报道和民众评论。对于司机，民众在微博评论中表现出的负面情感要高于正面情感，司机的行为引起网民的质疑和愤怒，"恐怖""故意"等负面情感词大量使用；后期随着对司机报复动机的调查，网民中出现"同情""平安"等正面的情感。

(3)主题 z_3 指的是关于车上"学生"和乘客的新闻报道和微博评论。从图中可以发现关于"学生"主题的情感表现正面明显高于负面，对于车上无辜学生，网民大多表示同情和希望平安，"幸运""平安""希望"等正面情感词被大量使用，负面情感主要是表现出的对无辜学生的"揪心""难受"等。

(4)主题 z_4 指的是关于"调查结果"的新闻报道和微博评论。从图中可以发现，该主题的情感表现负面高于正面，尤其是 7.12 以后负面情感明显高于正面，"难受""气愤""荒唐"等负面情感词被大量使用。

结束语　本文讨论了时间感知情感主题模型(TST)：一种新的基于主题模型的情感和主题动态联合建模方法。通过一个基于基本事实的评价框架，证明了 TST 在提取准确的

主题-情感-时间关联方面优于其他两个最先进的模型。同时还以突发事件“贵州公交车坠湖”微博数据集为例对 TST 模型进行了实证研究，发现模型具有很好的实际应用性，有助于分析突发事件中主题情感的演变，发现以往模型无法看到的隐藏现象，从而产生广泛的应用前景。

作为未来的发展方向，研究 TST 模型的超参数设置将是一个有趣的领域。图 2 和图 3 所示的结果表明，与其他主题情感模型相比，TST 对主题情感先验 γ 非常敏感，通常在几次实验后根据经验确定。目前正在进行的一项工作为检验文献中基于最大似然估计的方法[22]，结果表明这些方法在估计 α 和 β 超参数方面表现良好，但在估计 γ 方面表现较差。我们认为基于 TST 的主题情感分析过程的自动化应该基于用户的交互来指导分析过程。

参 考 文 献

[1] LI F, HUANG M, ZHU X. Sentiment analysis with global topics and local dependency[C]// AAAI'10. 2020: 1371-137.

[2] LIN C, HE Y, EVERSON R, et al. Weakly Supervised Joint Sentiment-Topic Detection from Text [J]. TKDE, 2012, 24(6): 1134-1145.

[3] XU Y M, LI Y, LIANG Y, et al. Topic-sentiment evolution over time: a manifold learning-based model for online news [J]. Journal of Intelligent Information Systems, 2020, 55: 27-49.

[4] KALARANI P, BRUNDA S S. Sentiment analysis by POS and joint sentiment topic features using SVM and ANN[J]. Soft Computing, 2019, 23: 7067-7079.

[5] HE Y L, LIN C H, GAO W, et al. Dynamic joint sentiment-topic model[J]. Acm Transactions on Intelligent Systems & Technology, 2014, 12(1): 1-21.

[6] GHOORCHIAN K, SAHLGREN M. GDTM: Graph-based Dynamic Topic Models [J]. Progress in Artificial Intelligence, 2020, 9: 195-207.

[7] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences, 2004(101): 5228-5235.

[8] CHANG C H, MONSELISE M, YANG C C. What Are People Concerned About During the Pandemic? Detecting Evolving Topics about COVID-19 from Twitter[J]. Journal of Healthcare Informatics Research, 2021, 5: 70-97.

[9] LI W B, MATSUKAWA T, SAIGO H. Context-Aware Latent Dirichlet Allocation for Topic Segmentation [J]. Advances in Knowledge Discovery and Data Mining, 2020(5): 475-486.

[10] HUANG L, TAN W N, SUN Y. Collaborative recommendation algorithm based on probabilistic matrix factorization in probabilistic latent semantic analysis [J]. Multimed Tools Appl, 2019, 78: 8711-8722.

[11] FATEMI M, SAFAYANI M. Joint sentiment/topic modeling on text data using a boosted restricted Boltzmann Machine [J]. Multimedia Tools and Applications, 2019, 78: 20637-20653.

[12] XU Y J, SUN C H, LIU Y Z. Joint sentiment/topic model integrating user characteristics[J]. Journal of Computer Applications, 2018(5): 1261-1266.

[13] JIANG C Q, LV X Z, DUAN R. Analyzing topic evolution of online product forum based on topic model[J]. Journal of Systems Engineering, 2019(10): 598-609.

[14] XU Y M, LI Y, LIANG Y, et al. Topic-sentiment evolution over time: a manifold learning-based model for online news[J]. Journal of Intelligent Information Systems, 2020, 55: 27-49.

[15] ZHU X X, SONG J X, MENG J F. Analysis of Online Public Opinion Information Based on the Dynamic Theme emotion Evolution Model[J]. Information Science, 2019(7): 72-78.

[16] HE Y, LIN C, GAO W, et al. Dynamic Joint Sentiment-Topic model[J]. TIST, 2014(9): 212-225.

[17] HE Y, LIN C, GAO W, et al. Tracking Sentiment and Topic Dynamics from Social Media[C]// ICWSM'12. Dublin, Ireland: AAAI, 2012: 483-486.

[18] WANG X, MOHANTY N, MCCALLUM A. Group and topic discovery from relations and text[C]// LinkKDD'05. Chicago, IL, USA: ACM, 2005: 28-35.

[19] HEINRICH G. Parameter estimation for text analysis[J]. Tech. Rep., 2005.

[20] SHI W, WANG H W, HE S Y. Study on Construction of Fuzzy Emotion Ontology Based on HowNet[J]. Journal of The China Society for Scientific Andtechnical Information, 2012(6): 595-602.

[21] BIGI B. Using Kullback-Leibler Distance for Text Categorization[C]// ECIR'03. Pisa, Italy: Springer-Verlag, 2003: 305-319.

[22] MINKA T P. Estimating a Dirichlet distribution[J]. MIT, Tech. Rep. 8, 2003.

SHI Wei, born in 1981, Ph.D, professor. His main research interests include business intelligence and affective computing.