doi: 10.13413/j.cnki.jdxblxb.2023475

基于动态主题情感模型的文本聚类算法

胡 萍
(铜仁学院 大数据学院, 贵州 铜仁 554300)

摘要: 针对目前已有的相关主题模型中, 对大众情感因素考虑不足, 难以精准挖掘, 同时对社交文本的实时动态演化考虑弱化了模型聚类能力的问题, 通过在模型中增加情感层以提取社交文本情感极性特征, 并引入先验分布函数, 提出一种基于动态主题情感模型的文本聚类算法. 利用真实新冠疫情 Twitter 文本数据集进行实验, 实验结果表明, 该模型的性能优于基线模型, 提高了情感特征区分度, 使文本主题与对应的情感极性联合生成时间节点, 进而使模型有处理时间演化的能力.

关键词: 动态主题情感模型; 文本挖掘; 情感标签; 时间戳; 文本聚类; 困惑度
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1671-5489(2025)02-0528-09

Text Clustering Algorithm Based on Dynamic Theme Emotion Model

HU Ping
(*School of Data Science, Tongren University, Tongren* 554300, *Guizhou Province, China*)

Abstract: Aiming at the problem that the emotional factors of the public were not considered enough in the existing related theme models, which was difficult to accurately excavate them, and the real-time dynamic evolution of social texts was considered to weaken the clustering ability of the model, the author proposed a text clustering algorithm based on the dynamic theme emotin model by adding the emotional layer to the model to extract the polar features of social text emotion, and introducing a prior distribution function. The experiments were carried out by using real COVID-19 Twitter text datasets. The experimental results show that the performance of the model is better than the baseline model, and the discrimination of emotional features is improved, so that the text theme and the corresponding emotional polarity can jointly generate time nodes, and then the model has the ability to deal with time evolution.

Keywords: dynamic topic emotion model; text mining; emotional label; time stamp; text clustering; perplexity

现有的主题模型主要包括情感主题模型和动态主题模型两种. 对于社交文本的情感主题模型, 目前的主要研究工作为主题情感建模[1-3], 主要特征为无监督学习, 但在情感特征分析方面存在不足, 且所使用的情感提取方法未考虑时间影响因子, 导致处理强时间性的社交文本表现不佳.

在动态模型研究中, 一部分在文本主题参数的时间演变中融入了 Markov 假设[4], 即当前主题是从上一时间节点的主题演化而来的, 词汇分布也随时间演化而动态变化. Liu 等[5]在数据集 Twitter 上

收稿日期: 2023-11-07.

作者简介: 胡 萍(1983—), 女, 土家族, 博士, 副教授, 从事数据挖掘和算法分析的研究, E-mail: 290222350@qq.com.

基金项目: 国家自然科学基金面上项目(批准号: 62066040)、教育部人文社科青年基金(批准号: 20YJC880030)和铜仁学院博士科研启动基金(批准号: trxyDH1914).

的实验表明，其提出的 FR-DATM 模型从生成的潜在主题的质量、模型的困惑度和可以动态挖掘作者关注主题三方面都优于潜在 Dirichlet 分配(latent Dirichlet allocation，LDA)和微博潜在 Dirichlet 分配(MB-LDA)模型. Ranganathan 等[6]提出了一种自动检测 Twitter 消息情绪的方法，该方法利用支持向量机 LibLinear 模型探索推文的特征和作者的情绪，并实现了 98%的准确率. 情感挖掘在计算机科学领域引起了广泛关注，因为它可以用于开发各种系统和具有潜力的应用，如远程医疗系统、客户服务、基于用户情感响应的智能手机以及感知驾驶员情感的车辆等. 这些情感有助于理解用户的当前状态，从而采取适当的行动或提供建议，以增强其对更健康生活方式的感知[7]. Zhang 等[8]设计并实现了基于时间序列新词挖掘的 COVID-19 舆情监测系统，提出了一种新的基于网络话题定时爆炸的词结构发现方案和针对 COVID-19 舆情环境的中文情感分析方法. 该系统可以根据评论判断评论者的积极情绪和消极情绪，也可以反映出希望、快乐、抑郁等 7 种情绪的深度. Xue 等[9]提出了一种结合情感标签和词间关系的语义情感主题模型(semantic emotion-topic model，SETM)，考虑了关联关系、计算时间、主题数和语义可解释性 4 个因素对 SETM 性能的影响.

目前，研究模型大多将时间演化因素纳入了考量范围，但未深度考虑文本的情感挖掘，特征区分度较低[10-12]. 基于此，本文提出一种 sToT 模型，以解决文本中情感特征区分度低的问题. 该模型在传统主题模型的基础上增加了情感层，以提取文本中主题的情感极性特征，增强了情感区分度. 为捕捉主题的动态演化关系，在模型中引入时间先验分布函数，使主题和对应的情感极性有机融合，进而观察到时间戳和单词，使模型建模凸显出时间因素对主题的演化影响. 并选取新冠疫情 Twitter 文本数据集进行实验，实验结果表明，该模型展现了优于基线模型的性能，验证了该模型的有效性.

1　动态主题情感模型

本文使用的输入数据为新冠疫情期间社交媒体 Twitter 中的文本数据. 在使用前，首先对 Twitter 的文本数据进行分词、去停词、提取词干、还原词干等操作，然后在原有模型的基础上加入情感词典，以对单词进行情感打分，进而提升文本的情感区分度. 为使文本主题与时序演化相对应，本文模型引入时间先验分布函数以进行主题情感数据的动态挖掘[13-15]. 在模型的推理部分，根据单词、主题、情感以及时间戳联合概率公式对算法使用 Gibbs 采样. 模型迭代过程中会为单词分配各自的主题号、情感标签以及时间戳直至收敛到稳定状态，然后输出文本的情感强度值，直观体现文本主题的情感子空间演变及文本单词的聚类结果.

1.1　任　务

本文模型选取的文本是以新冠肺炎为关键词的社交媒体文档集 $D=\{d_1,d_2,\cdots,d_D\}$，该文档集中包含社交媒体用户所发送的文本内容及社交媒体形成的时间戳 $V=\{W_i,t_i\}$，其中 W 为文档集 D 中所有单词组成的集合，各文档中的单词都是表 $V=\{1,2,\cdots,N\}$中的索引，t 为社交文档生成时的时间戳. 在对社交文本进行建模时，各主题都有对应的两种情感极性，对这些文档进行分析时，要分析相应的主题.

本文模型使用了 Sentiwordhel 的情感词标注功能. 通过为文本中的单词进行情感标签分配，并赋予相应的情感分值，再综合单词、主题、情感和时间截的联合概率进行 Gibbs 采样，通过这种形式进行主题号、情感标签和时间戳的分配，为下次迭代使用提供依据. 在模型迭代至概率分布均达到收敛稳定状态时，计算单词的条件概率分布以主题情感词列表的形式表示文本聚类结果.

1.2　模　型

下面对主题情感动态演化过程及模型中的主题情感关联方法进行阐述，并利用该模型挖掘出社交文本中潜在的情感极性变化情况. 本文提出的 sToT 模型如图 1 所示，其中 α 为主题先验参数，β 为主题情感词先验参数，φ 为时间先验参数，D 为文档总数，N_d 为每篇文档 d 中的词汇数量，θ 为主题分布，z 为主题号，t 为时间截，φ 为主题情感词分布，T 为主题数，S 为情感数，w 为词汇数. 图 1 中的矩形表示重复过程. 其中的观测变量包括情感极性和单词项，箭头表示各参数之间的依赖关系，学习数据中的文档必须标注创建时间. 在时间标注方面，将日期转换为 Unix 时间戳格式并进行归一化，最

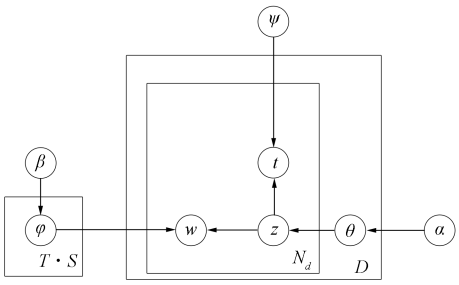

图1 sToT模型

Fig. 1 SToT model

后为每个文档中的单词进行标注. 由图1可见，在文档生成过程中，先根据主题参数得到每篇文档的主题分布，然后根据主题-情感得到情感词，再通过外部的情感字典，对每个取出的情感词标注情感标签(正或负)，如果该单词不在情感词典中，则随机选取一个标签，对文档中的主题和情感进行混合分布. 此外，在迭代过程中，用变量 t 捕获时间截，使用时间戳 t 上的 ψ 多项分布获取主题情感的演化过程.

在建模过程中本文未对时序状态转换进行离散化，避免了Markov假设对时间离散化的影响. sToT模型对每个主题与情感相关的连续时间分布都进行了参数化，使得每个主题与其对应的情感在时间区间内实现了连续性的高度关联. 因此，主题标签和情感标签联合生成时间戳和单词，使主题和情感的生成不仅受单词的影响，也受时间戳的影响.

在文本动态建模过程中，本文使用Beta分布对时间戳进行采样，在[0,1]内进行连续性概率分布，公式如下：

$$f(x;\alpha,\beta)=\frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{1}$$

其中参数 $\alpha,\beta>0$. 将参数的时间戳范围归一化到[0,1]区间，以实现其在时间轴上的前后关联. 通过将主题及其对应的情感随时间变化设为常数，本文观察到了主题与情感本身的演变过程，而非主题情感词分布的变化. 同时，本文还将主题与情感定义为动态共存的关系. 在构建情感模型时，引入了情感层扩展现有的主题模型，并利用时间先验函数分析文本情感与时间的演化关系.

本文以传统主题模型为基础进行建模，迭代过程中所有文档都遵循多项式分布原则，由多个主题混合组成. 此外，假设每个主题的情感极性包括消极属性和积极属性两种类型. 因此，主题与情感的关联强度会随时间的推移而发生变化. 通过引入情感层，可以扩展现有的主题模型，并使用时间先验函数分析文本情感与时间的演化关系. 在构建情感模型时还需考虑每个主题的情感极性如何随时间变化，以及主题与情感之间的动态共存关系.

1.3 情感挖掘模块

本文使用的情感挖掘方法参考了文献[1]提出的模型，模型结构如图2所示，其中 γ 为情感先验参数，π 为每个主题的情感分布，l 为情感标签，其他符号的含义同图1.

本文情感模块在进行情感特征挖掘时引入了外部的情感词典，为模型中的单词打上极性标签. 此外，本文情感模型在建模时，基于单词在不同主题下有不同的极性，所以假设主题决定情感的极性，以建立主题与情感之间的关联.

本文情感挖掘模块能构建情感层与主题层之间的联系，将单词的生成与情感标签、主题标签相关联，对各主题的情感极性进行分析. 情感标注上，引入了外部的情感词典，在情感极性上也只设置了消极与积极两个因子，积极的情感极性标签 l 的值为1，消极为0. 为使模型算法能收敛，其计算条件概率公式如下：

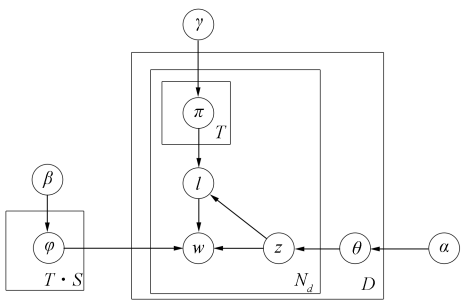

图2 情感挖掘模块

Fig. 2 Module of emotion mining

$$p(z_i=z,\ l_i=l\,|\,z_{i\neg},l_{i\neg},w)\propto\frac{\{n_m^{(z)}\}_{i\neg}+\alpha}{\{n_d\}_{i\neg}+T\alpha}\cdot\frac{\{n_d^{(z,l)}\}_{i\neg}+\gamma_l}{\{n_d^{(z)}\}_{i\neg}+\sum_{l=1}^{S}\gamma_l}\cdot\frac{\{n_{z,l}^{(t)}\}_{i\neg}+\beta}{\{n_{z,l}\}_{i\neg}+V\beta}, \tag{2}$$

其中 w 表示单词，γ_l 表示情感 l 的先验参数，$n_d^{(z)}$ 表示第 d 篇文档中指定给主题 z 的单词数，n_d 表示

第 d 篇文档中的总单词数，$n_d^{(z,l)}$ 表示第 d 篇文档中指定给主题 z 和情感 l 的单词次数，$n_{z,l}^{(t)}$ 表示单词出现在主题 z 和情感 l 中的次数，$n_{z,l}$ 表示主题 z 和情感 l 下的单词数，$\neg$ 表示除去第 d 篇文档中第 i 个单词的计数次数.

在情感标签生成方面，使用 NLTK 中的 Sentiwordnet 情感词典. 词典示例列于表 1.

表 1　Sentiwordnet 示例

Table 1　Example of Sentiwordnet

词性	ID	Pos	Neg	SynsetTerms # sentimentscore	Gloss
a	00009618	0.5	0.25	Spartan # 4	describe

在 SentiWordNet 中，每个单词都包含词性信息，其中 n 表示名词，a 表示形容词，v 表示动词，r 表示副词. 此外，每个单词还具有词条编号以及积极情感得分和消极情感得分. 对于同义词，单词之间存在相同的词条编号. 在注释部分，可以发现单词的其他信息. 但在实际文本数据中，一个单词可能具有多种含义. 例如“good”作为名词有 4 种含义，而作为形容词则有 21 种含义，作为副词有 2 种含义. 在本文实验中，对每个单词进行了积极情感和消极情感的标注，标注过程考虑了单词的多种含义和词性，以确保对情感倾向的准确评估，计算公式分别为

$$\mathrm{posScore}_{\mathrm{word}}=\sum_{i=1}^{n}\frac{\mathrm{Pos}_i}{n}, \tag{3}$$

$$\mathrm{negScore}_{\mathrm{word}}=\sum_{i=1}^{n}\frac{\mathrm{Neg}_i}{n}, \tag{4}$$

其中 n 表示单词所有含义的数量，Pos_i 表示单词当前第 i 个含义中的积极情感得分，Neg_i 表示单词当前第 i 个含义中的消极情感得分. 若 $\mathrm{posScore}_{\mathrm{word}}\geqslant 0.1$ 且 $\mathrm{posScore}_{\mathrm{word}}>\mathrm{negScore}_{\mathrm{word}}$，则给单词赋子情感标签 1，表示积极情感标签，若 $\mathrm{negScore}_{\mathrm{word}}\geqslant 0.1$ 且 $\mathrm{negScore}_{\mathrm{word}}>\mathrm{posScore}_{\mathrm{word}}$，则给单词赋予消极情感标签 0，表示消极情感标签.

1.4　动态挖掘模块

sToT 模型中的动态挖掘模块整体框架如图 3 所示. 与采用 Markov 假设建模时间序列不同，该模块每个主题都与时间戳上的连续分布相关联. 此外，对每个生成的文档，其主题分布都受文档时间戳的影响. 因此，该模块能挖掘主题随时间变化的规律.

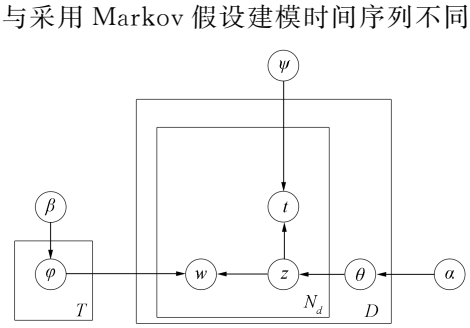

图 3　动态挖掘模块

Fig. 3　Dynsmic mining module

该模型架构使主题发现不仅受单词共现影响，也受时间信息影响. 在建模时会对时间戳进行归一化，而不是采用动态 Markov 假设建模状态变化序列. 这样可以使模型在时间上观察到长期的依赖关系，也有助于避免 Markov 模型的风险，即在主题客观上存在短暂间隙时错误地将其分为两个主题. 模型的生成过程如下：

1）采样出主题分布 $\varphi_z\sim\mathrm{Dirchlet}(\beta)$；

2）对于每篇文档，采样出一个多项式分布 $\theta_d\sim\mathrm{Dirchlet}(\alpha)$；

3）对文档中每个单词，采样一个主题 $z_{di}\sim\mathrm{Mulitnomial}(\theta_d)$，采样一个单词 $w_{di}\sim\mathrm{Mulitnomial}(\varphi_{z_{di}})$，采样一个时间数 $t_{di}\sim\mathrm{Beta}(\psi_{z_{di}})$.

模型采用 Gibbs 采样进行近似推理，需要计算条件概率为

$$p(z_{di}\mid w,t,z_{di},\alpha,\beta,\psi)\propto(m_{dz}+\alpha_{z_{di}}-1)\cdot\frac{n_{z_{di}w_{di}}+\beta_{w_{di}}-1}{\sum_{v=1}^{V}(n_{z_{di}}+\beta V)-1}\cdot\frac{(1-t_{di})^{\hat{\psi}^1_{z_{di}}}t_{di}\hat{\psi}^2_{z_{di}}-1}{\mathrm{B}(\hat{\psi}^1_{z_{di}},\hat{\psi}^2_{z_{di}})}, \tag{5}$$

其中：t_{di} 表示第 d 篇文档中第 i 个单词的时间截；V 表示词典大小；w_{di} 表示第 d 篇文档中第 i 个单词；

$n_{z_{di}v}$表示第d篇文档中第i个单词v分配给主题z的次数；m_{dz}表示文档d中主题z出现的次数；$\hat{\psi}^1_{z_{di}}$，$\hat{\psi}^2_{z_{di}}$表示用于时间建模的Beta分布的两个参数，通过矩量法估计. 在时间建模部分，上述过程中每个单词都与一个时间戳相关联. 在训练数据拟合模型时，每个训练文档中的所有单词都将被赋予相同的时间戳.

2 动态主题情感模型的生成过程

2.1 参数估计

sToT模型是一个面向带时间戳的文档中的单词的生成模型，其通过Gibbs采样过程对参数进行估计，生成过程如下：

1) 从先验参数β中采样出维度为$T\times S$的主题z情感l下的单词分布$\varphi_{zl}\sim$Dirchlet(β)；

2) 对每篇文档d，从先验参数a中采样得到一个主题的多项式分布$\theta_d\sim$Dirchlet(α)；

3) 对于文档中的每个单词w_{di}：采样生成一个主题$z\sim$Mulitnomial(θ_d)，采样生成文档中每个主题下的情感分布$\pi_{dz}\sim$Dirchlet(γ)，采样生成情感标签$l\sim$Bernoulli(π_{dz})，采样生成单词$w_a\sim$Multnomial($4z$)，采样生成时间戳$w_{di}\sim$Multnomial(φ_{zl})，采样生成时间戳$t_{di}\sim$Beta(ψ_{zl}).

重复步骤d次，则生成一篇文档，整体步骤重复D次，则生成一个文档集. 根据sToT模型的生成过程及图模型可见，同一文档中的不同单词可能会生成不同的时间戳. 但在文档中的所有单词应该具有相同的时间戳. 主题模型的推理过程实际上是文档生成过程的逆向推理. 在实际应用中，很难精确求得概率分布，此时需通过概率统计推导获得文档的隐含信息. 本文根据文档的可观测变量进行逆向推理，得到主题-情感分布和主题-词分布. 通常情况下，主题模型采用近似推断方法，主要有Gibbs采样和变分推断两种. 本文采用Gibbs采样算法.

2.2 模型推导

Gibbs采样是主题模型参数估计中的一种流行方法，其通过迭代的方式对复杂的概率统计问题进行求解推导. 当模型经过足够多次迭代后，将达到收敛稳定的状态，此时再进行迭代就不会产生较大变化. 在稳定状态下得到的主题-情感分布和主题-词分布将最接近于文档的真实分布. 本文通过使用Bayes条件概率公式求解联合概率分布. 其中，单词、主题、情感和时间戳的联合概率可分解为

$$p(w,t,l,z\mid\alpha,\beta,\gamma,\mu)=p(w\mid l,z,\beta)\cdot p(l\mid z,\gamma\mid)\cdot p(z\mid\alpha)\cdot p(t\mid l,z,\psi). \tag{6}$$

通过对式(6)中第一项中的隐变量φ积分，整理后可得

$$p(w\mid l,z,\beta)=\int p(w\mid z,l,\varphi)\cdot p(\varphi\mid\beta)\mathrm{d}\varphi=\left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{T\cdot S}\prod_j\prod_k\frac{\prod_i\Gamma(n_{i,j,k}+\beta)}{\Gamma(n_{j,k}+V\beta)}, \tag{7}$$

其中Γ表示Gamma函数，i,j,k分别用于循环单词、主题、情感. 以此类推，第二项整理后可得

$$p(l\mid z,\gamma)=\int p(l\mid z,\pi)\cdot p(\pi\mid\gamma)\mathrm{d}\pi=\left(\frac{\Gamma\left(\sum_k r_k\right)}{\prod_k\Gamma(r_k)}\right)^{T\cdot D}\prod_j\prod_d\frac{\prod_k\Gamma(n_{j,d,k}+r_k)}{\Gamma\left(n_{j,d}+\sum_k r_k\right)}. \tag{8}$$

第三项整理后可得

$$p(z\mid\alpha)=\int p(z\mid\theta)\cdot p(\theta\mid\alpha)\mathrm{d}\theta=\left(\frac{\Gamma(T\cdot\alpha)}{\Gamma(\alpha)^T}\right)^D\prod_d\frac{\prod_j\Gamma(n_{d,j}+\alpha)}{\Gamma(n_d+T\cdot\alpha)}. \tag{9}$$

在动态建模中，本文未使用离散化时间，而是采用了连续的Beta分布. 由于时间戳来自连续的Beta分布，所以对ψ参数的估计，稀疏性不是一个重要问题. 为简化和加快计算，本文在每次Gibbs采样迭代中都通过矩量法更新Beta分布ψ，用公式表示为

$$\hat{\psi}^1_{zk}=\bar{t}_{zk}\cdot\left(\frac{\bar{t}_{zk}(1-\bar{t}_{zk})}{s^2_{zk}}-1\right), \tag{10}$$

$$\hat{\psi}^2_{zk}=(1-\bar{t}_{zk})\cdot\left(\frac{\bar{t}_{zk}(1-\bar{t}_{zk})}{s^2_{zk}}-1\right), \tag{11}$$

其中 z,s 分别表示属于主题 z 和情感 k 的时间戳的样本均值和样本方差. 整理后并结合式(1)，可推导出第四项公式为

$$p(t\mid l,z,\psi)=\prod_{d}\prod_{i}\prod_{k}p(t_{di}\mid\psi_{zk})=\prod_{d}\prod_{i}\prod_{k}\frac{(1-t_{di})^{\hat{\psi}^1_{zk}-1}\cdot t_{di}^{\hat{\psi}^2_{zk}-1}}{\mathrm{B}(\hat{\psi}^1_{zk},\hat{\psi}^2_{zk})}. \tag{12}$$

综上，通过给定所有其他变量，对主题和情感采样计算后验概率分布，用 p 表示当前文档除位置 p 处单词的统计次数，利用式(6)推出后验概率为

$$p(l_p=k,\ z_p=j\mid w,t,l_{p\neg},z_{p\neg},\alpha,\beta,\gamma,\psi)=\frac{p(w,t,l,z\mid\alpha,\beta,\gamma,\psi)}{p(w_{p\neg},t_{p\neg},l_{p\neg},z_{p\neg}\mid\alpha,\beta,\gamma,\psi)}\propto$$

$$\frac{n_{i,j,k}+\beta}{n_{j,k}+V\beta}\cdot\frac{n_{j,d,k}+r_k}{n_{j,d}+\sum\limits_{k}r_k}\cdot\frac{n_{d,j}+\alpha}{n_d+T\cdot\alpha}\cdot\frac{(1-t_{d,i})^{\hat{\psi}^1_{zk}-1}\cdot t_{d,i}^{\hat{\psi}^2_{zk}}-1}{\mathrm{B}(\hat{\psi}^1_{zk},\hat{\psi}^2_{zk})}, \tag{13}$$

其中 $n_{i,j,k}$ 表示单词 i 分配给主题 j 和情感 k 的次数，$n_{d,i}$ 表示分配给主题 j 情感 k 的次数，$n_{i,d,k}$ 表示文档 d 中分配给主题 j 和情感 k 的次数，$n_{d,i}$ 表示文档 d 中分配给主题 j 的次数，r_k 为情感 k 对应的先验参数. 根据上述更新规则，经过一系列 Gibbs 采样后，可根据得到的参数近似计算主题分布 β、主题情感词分布 φ 和主题情感分布 π，用公式分别表示为

$$\theta_{d,j}=\frac{n_{d,j}+\alpha}{n_d+T\cdot\alpha}, \tag{14}$$

$$\varphi_{i,j,k}=\frac{n_{i,j,k}+\beta}{n_{j,k}+V\cdot\beta}, \tag{15}$$

$$\pi_{d,j,k}=\frac{n_{i,j,k}+\gamma_k}{n_{j,d}+S\cdot\gamma}. \tag{16}$$

3　仿真实验与结果分析

本文实验中采用了 Python 语言和 Numpy 开源的数据计算模块. 为验证模型的有效性，在定量分析方面进行多组对照实验，以证明本文设计的模型框架相对于其他模型在性能方面具有一定的优越性. 在定性分析方面，通过可视化展示结果，直观显示了主题及其对应情感随时间的演变趋势，并结合实际生活进行分析.

3.1　数据集及模型评估指标

3.1.1　数据集获取

本文实验所用的数据来自 Kaggle 平台中的英文新冠疫情 Twitter 数据集，以设定关键词如 COVIDUSA，COVID19，Coronavinus，SARSCoV2SocialDistance，washhands，safehandsQuarantineLife 在 Twitter 平台上进行内容搜索，结果包含了用户发表的帖子和帖子建立的时间. 实验共选择 29 585 条文本，时间跨度为 2020-03-03—2020-04-29.

对爬取原始文本数据进行遍历构建停词表去停词、词性还原、删除标点符号并进行分词操作. 经过上述操作的英文文本预处理和整合后，英文 Twitter 数据集部分示例列于表 2.

表 2　英文 Twitter 数据集示例

Table 2　Example of English Twitter dataset

文本	时间
hey pandemic your get heart lockdown	2020-04-29 23:52
covid teach anything nearly much control think make plan lord determine step	2020-04-29 23:53

3.1.2　模型评估指标

为评估模型性能，本文主要使用困惑度(Perplexity)作为评价指标. 在信息论的测量中，困惑度被用来度量一个概率分布或概率模型对样本的预测能力，用于比较不同主题模型的性能. 模型困惑度越低，表示主题模型性能越好. 困惑度的计算公式为

$$\mathrm{Perplexity}(D)=\exp\Big\{-\Big(\sum_{w=1}^{n}\log\ p(w)\Big)\Big/\sum_{d=1}^{M}N_d\Big\}, \tag{17}$$

其中 D 表示文档集合，M 表示文档总数，$p(w)$表示文档中单词出现的概率，N_d 表示第 d 篇文档中的单词数.

3.2 模型的评估

3.2.1 定量分析

主题情感模型的超参数值选择十分重要. 在模型性能验证对比实验中，采用常用的参数设置方法，包括主题先验参数、主题情感词先验分布参数和情感先验分布参数. 实验主要对比本文模型与如下几种当前挖掘主题情感的代表性主题模型的性能.

1) LDAM：使用潜在 Dirichlet 分布的概率模型，起到推测文档主题分布的作用，通过概率分布的方式呈现出文档主题，从而进行主题聚类.

2) JSTP：一个基于潜在 Dirichlet 分布的概率模型，用于探讨文档主题情感的模型，对文本中的主题和情感进行检测，研究文档级情感分类的主题情感模型. 模型中，情感标签与文档相关联，主题与情感标签相关联，单词与情感标签和主题两者相关联.

3) Sentiment-LDA：一个假设情感与主题相关的主题情感模型，模型中一个单词的情感特征取决于所属的主题，不仅可以对文档的整体情感进行分类，还可以计算每个主题的情感特征.

实验中，本文比较了不同主题数对不同模型性能的影响. 设主题数分别为 50，60，70，80，90 和 100. 图 4 为本文模型与不同基线模型在不同主题数下的困惑度数值比较. 由图 4 可见，本文模型性能最好. 导致模型性能差异的主要原因是基线模型仅考虑了文本中的情感信息或主题，而忽视了文本的时间属性. 因此，这些方法在建模过程中存在较大的困惑度. 而本文模型在主题情感模型的基础上，利用 Beta 分布对所有数据的时间戳信息进行建模，使模型能更好地捕捉 Twitter 文本中的主题情感演化过程，并进一步提高了模型性能.

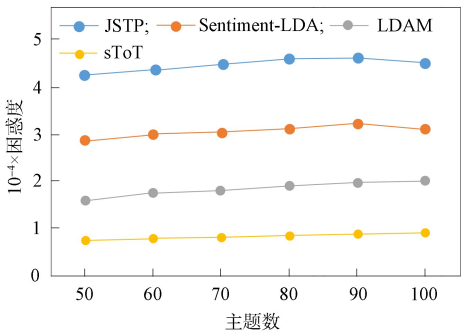

图 4 不同主题数下不同模型的困惑度比较

Fig. 4 Comparison of perplexity of different models under different theme numbers

3.2.2 定性分析

为进一步验证主题情感模型的有效性，下面从时间演化视角以及主题情感挖掘两方面进行验证. 定性分析结果显示，结合概率分布可以展示排名靠前的情感倾向、情感词和主题词. 设主题数量为 50 个，表 3 列出了从数据集中提取的部分主题情感词示例. 由表 3 可见，本文模型能很好地从文本数据中挖掘出主题及其情感特征. 其中许多热点主题，如戴口罩、居家令以及疫情期间人们的恐慌等，都得到很好地提取.

表 3 主题情感词示例

Table 3 Example of theme emotion words

z1：隔离		z2：封城		z3：选举		z4：商店关门	
积极	消极	积极	消极	积极	消极	积极	消极
safeguard	shutdown	prevention and control	lockdown	democracy	fraud	reopening	closure
care	quarantine	protection	epidemic	freedom	violence	transformation	bankruptcy
support	loneliness	health	infection	equality	corruption	upgrading	unemployment
cooperation	boredom	safety	death	fairness	manipulation	innovation	business interruption
resilience	anxiety	unity	isolation	participation	ntimidation	adaptation	loss
solidarity	frustration	confidence	restriction	responsibility	warn	adjustment	difficulty

经过可视化后，图 5 显示了表 3 中 4 种主题-情感随时间的演变过程.

结合表 3 中主题情感词聚类示例结果可得以下信息：

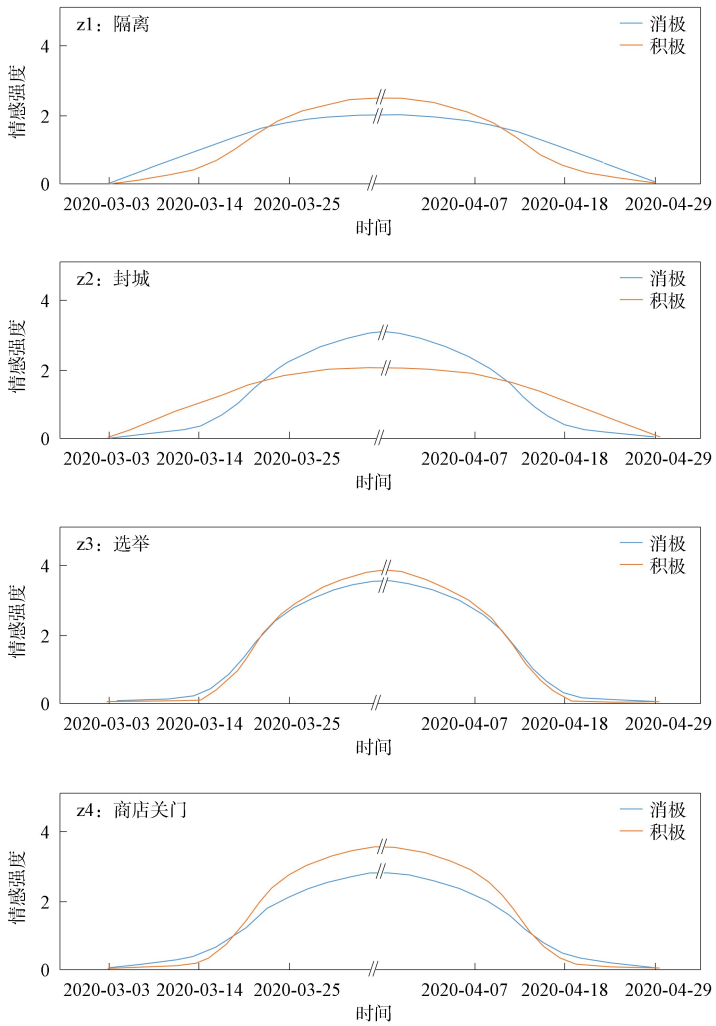

图 5　主题-情感演化示例

Fig. 5　Examples of evolution of themes and emotions

1）主题 z1 描述了在颁布居家隔离政策后，人们的心理活动变化. 消极情感中出现了 quaranti，challenge 等词，可得出隔离是一种挑战，而积极情感中出现 stayathome，music，love，listen，study，可得出隔离可以听音乐，也可以学习，并且人们十分享受这样的状态. 由图 5 可见，消极情感增长明显比积极情感快，并在 3 月底达到高峰，远高于积极情感，说明人们随着时间变化对居家隔离政策呈抵触态度.

2）主题 z2 描述了人们对封城令的情感态度. 积极情感中出现了 protect，reduce 单词，得出封城有助于保护民众健康，减少了疫情的扩散. 而消极情感中出现了 old，scari 等词，得出封城政策使民众感到害怕. 由图 5 还可见，两者增长趋势基本一致，但在 3 月底时消极情绪显然明显高于积极情绪，因此反应出人们更多的还是恐惧和害怕.

3）主题 z3 描述了疫情期间美国政府对 2020 年大选的态度，积极情感中出现 right，boost，belif，得出对于大选持有支持态度，而消极情感中出现 warn，bad，improper，得出这类情感认为大选不合时宜且十分糟糕. 由图 5 可见，两种情绪无论是增长态势还是高峰期都类似，反应出人们对疫情下的美国大选无论是支持还是反对都显现出对峙的形式.

4）主题 z4 描述了人们对疫情中商店关门的情感. 积极情感中出现 love，survive，killcovid 单词，可以得出民众对遏制新冠和存活的渴望，消极情感中出现 march，pinch，得出人们对商店关门的抵触. 总体而言，积极情感强度高于消极情感强度，说明更多的人还是希望商店关门、减少流动，从而缓和疫情.

综上所述，针对目前已有的相关主题模型中，对大众情感因素考虑不足，难以精准挖掘，同时对社交文本的实时动态演化考虑弱化了模型聚类能力的问题，本文提出了一种基于动态主题情感的文本聚类模型，该模型在传统主题模型的基础上通过增加情感层提取文本中主题的情感极性特征，增强了文本特征的情感区分度. 此外，引入了时间先验分布函数，使主题和对应的情感极性联合生成观察到的时间戳和单词，使模型建模了时间变化对主题的影响. 在真实新冠疫情 Twitter 文本数据集上的实验结果表明，本文模型性能优于基线模型，从而验证了本文模型的有效性.

参 考 文 献

[1] LI Y H, FENG L Q. Opinion Mining for Multiple Types of Emotion-Embedded Products/Services through Evolutionary Strategy [J]. Expert Systems with Application, 2018, 61(4): 1874-1883.

[2] AVCI U. A Pattern Mining Approach for Improving Speech Emotion Recognition [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2022, 23(4): 37-45.

[3] ARYA A, SHUKLA V, NEGI A, et al. A Review: Sentiment Analysis and Opinion Mining [J]. International Journal of Research in Engineering and Applied Science, 2016, 6(10): 16-21.

[4] GU Z Y, LIN Y, DAI Y H, et al. Research on Online Emotion of COVID-19 Based on Text Sentiment Analysis [J]. International Journal of Computational Science and Engineering, 2022, 25(4): 460-466.

[5] LIU B Y, WANG C R, WANG Y R, et al. Microblog Topic Mining Based on FR-DATM [J]. Chinese Journal of Electronics, 2018, 12(9): 241-246.

[6] RANGANATHAN J, TZACHEVA A. Emotion Mining in Social Media Data [J]. Procedia Computer Science, 2019, 159(9): 58-66.

[7] LUNA D S, BERING J M. Varieties of Awe in Science Communication: Reflexive Thematic Analysis of Practitioners' Experiences and Uses of This Emotion [J]. Science Communication, 2022, 44(3): 347-374.

[8] ZHANG Y X, CHEN J R, LIU B Y, et al. COVID-19 Public Opinion and Emotion Monitoring System Based on Time Series Thermal New Word Mining [J]. Computers, Materials & Continua, 2020, 64(3): 1415-1434.

[9] XUE R R, HUANG S, LUO X, et al. Semantic Emotion-Topic Model Based Social Emotion Mining [J]. Journal of Web Engineering, 2018, 17: 73-92.

[10] MATHEW M K, SURYA R, ROSHAN J O, et al. Emotion Recognition Systems and Emotion Correlation Mining [J]. International Journal of Engineering Research & Technology, 2021, 9(7): 24-28.

[11] VINLUAN A, GONEDA M, ATIENZA F A L, et al. Opinion to Emotion Mining: A Sentiment Analysis towards Super Typhoon Ompong [J]. SSRN Electronic Journal, 2021, 11(1): 20-31.

[12] CASILLAS L A, ALEJANDRO R. Emotion Mining Mechanism over Texts in Social Media [J]. Research in Computing Science, 2019, 148(7): 227-240.

[13] RAÚL O B, RAMÓN Z C, LUCÍA B E M, et al. Opinion Mining and Emotion Recognition in an Intelligent Learning Environment [J]. Computer Applications in Engineering Education, 2019, 27(1): 90-101.

[14] BHAGATG V P. Emotion Extraction Using Ensemble Classification Model in Data Mining [J]. International, Organization of Scientific Research, 2018, 8(12): 14-22.

[15] PLAZA-DEL-ARCO F M. Lexicon Adaptation for Spanish Emotion Mining [J]. Procesamiento de Lenguaje Natural, 2018, 7(5): 661-671.

(责任编辑：韩 啸)