

文章编号: 1003-0077(2020)07-0096-09

基于性格情绪特征的改进主题情感模型

李玉强¹, 黄瑜¹, 孙念¹, 李琳¹, 刘爱华²

(1. 武汉理工大学 计算机科学与技术学院, 湖北 武汉 430063;

2. 武汉理工大学 能源与动力工程学院, 湖北 武汉 430063)

摘要: 近年来,以微博为代表的社交媒体在情感分析中备受关注。然而,绝大多数现有的主题情感模型并没有充分考虑到用户性格特征,导致情感分析结果难尽人意。故该文在现有的 JST 模型基础上进行改进,提出一种基于时间的性格建模方法,将用户性格特征纳入主题情感模型中;鉴于微博数据包含大量的表情符号之类的特有信息,为了充分利用表情符号来提升微博情感识别性能,该文将情感符号融入 JST 模型中,进而提出了一种改进的主题情感联合模型 UC-JST(Joint Sentiment/Topic Model Based on User Character)。通过在真实的新浪微博数据集上进行实验,结果表明 UC-JST 情感分类效果优于 JST、TUS-LDA、JUST、TSMFM 四种典型的无监督情感分类方法。

关键词: 主题情感模型;时间;性格特征;表情符号

中图分类号: TP391

文献标识码: A

An Improved Topic Sentiment Model Based on User Character

LI Yuqiang¹, HUANG Yu¹, SUN Nian¹, LI Lin¹, LIU Aihua²

(1. School of Computer Science and Technology, Wuhan University
of Technology, Wuhan, Hubei 430063, China;

2. School of Energy and Power Engineering, Wuhan University of
Technology, Wuhan, Hubei 430063, China)

Abstract: In the sentiment analysis in micro-blogs, most existing topic sentiment models do not fully consider the user's personality characteristics. Based on the JST model, this paper proposes a time-based personality modeling method to incorporate user's personality features into the topic sentiment model. Since the microblog data contains a lot of unique information such as emoticons, we also introduce emoticons into the JST model. As a result, an probabilistic model named UC-JST(Joint Sentiment/Topic model based on User Character)is proposed. Tested on the real Sina Weibo dataset, the results show that UC-JST performs better than JST, TUS-LDA, JUST and TSMFM in terms of sentiment classification accuracy.

Keywords: topic sentiment model; time; personality features; emoticons

0 引言

第 44 次《中国互联网络发展状况统计报告》数据显示,截至 2019 年 6 月,中国互联网用户规模达到 8.54 亿,互联网普及率为 61.2%。越来越多的用户习惯在微博平台上表达自己对某个事件的看法,而在信息迅猛扩散的时代,这些微博信息已成为挖

掘人们观点与情感的重要资源,为用户满意度调查、舆情监测、社会学研究等应用提供有效的数据支持^[1],因而对微博进行情感分析具有十分重要的现实意义。

微博情感分析的巨大价值促进了各种各样微博情感分析方法的产生,包括有监督的情感分类方法和无监督的情感分类方法。有监督的情感分类方法主要是基于传统的机器学习算法,该类算法需要标

收稿日期: 2019-09-26 定稿日期: 2019-11-18

基金项目: 国家社会科学基金(15BGL048)

注大量训练样本,而现实中已标注样本通常较少,使得有监督的情感分类方法需要付出很高的标注代价。而无监督情感分类方法克服了上述缺点,不需要标记任何样本,节省了高额的人工成本,当前基于主题模型的方法已经成为最广泛使用的无监督情感分类方法。

主题情感联合模型 JST (Joint Sentiment/Topic Model)^[2]是一种能够有效捕捉文档中隐含主题和情感的非监督学习方法,其同时考虑了单词的主题和情感,在 LDA (Latent Dirichlet Allocation) 模型基础上加入了情感层,可以对文档中不同主题与情感中的单词分布进行分析,取得了较好的情感分析效果。然而微博文本中富含大量的表情符号,有时一个表情符号可以完全改变文本的情感类别,若在进行微博情感分析时,将表情符号视为噪声数据而抛弃,很可能导致微博情感极性的误判。而 JST 模型只考虑文本中单词出现的情况,忽视了表情符号,导致 JST 模型在微博情感分析方面取得的效果并不突出。另外,人工心理学相关研究成果也表明,不同性格的人对同样的事物产生的情感反应是不同的。性格乐观者往往会通过不同方式与不同媒介来表达自己的积极的情感与态度,而性格悲观者则相反^[3]。由此可见,性格情绪特征对微博情感分类也能产生不小的影响,通过对性格的建模,可以对情感分析的效果产生一定的促进作用。

针对以上问题,本文提出将微博文本中的表情符号和用户性格情绪特征与 JST 模型相结合,弥补 JST 模型中微博特征缺失的问题,从而建立起了 UC-JST 模型。

1 相关工作

基于主题模型的方法主要分为以下两类:PLSA (Probability Latent Semantic Analysis) 和 LDA (Latent Dirichlet Allocation)^[4]。PLSA 引入潜在的主题层,将文档和词语都看作是随机变量,解决了 LSA 模型中一词多义的问题,但是会出现过拟合现象。LDA 在 PLSA 模型的基础上,为每一个主题增加了一个狄利克雷的先验分布,较好地解决了 PLSA 模型的过拟合问题,也逐渐成为主题模型中情感分析的主流方法。不少学者在 LDA 模型基础上进行了一定的改进:蔡永明等^[5]在传统 LDA 模型的基础上构建词汇社交网络,同时在吉布斯采样过程中增加隐含位置聚类模型的社区发现算

法,使得该模型在短文本主题分析中取得了较好的效果。Nguyen 等人^[6]在维基百科语料库上通过 Word2Vec^[7]训练出词向量,以改进 LDA 主题模型中的主题—词语映射,通过外部语料信息进一步丰富单词上下文语义,提高了模型的主题挖掘能力,但是这种词嵌入主题模型在大型语料库上训练速度太慢。

由于 LDA 模型在各个领域得到了广泛的应用,于是有学者尝试利用主题模型辅助情感分析,构造主题情感的混合模型。Lin 等人^[2]提出一种基于 LDA 模型的 JST 模型,真正实现将情感与主题联合建模,该模型加入情感层,形成一个包含词、主题、情感和文档的四层贝叶斯概率模型。JST 模型不仅可以产出高质量的情感相关的主题,而且在文档的情感分类的任务上也具有较好的表现。而后,他们结合情感词典 MPQA (Multi-Perspective Question Answering) 作为先验信息纳入 JST 模型中,并提出了 Reverse-JST 模型^[8]。在 Reverse-JST 中,情感与文档主题相关联,这意味着 Reverse-JST 考虑对主题级别的情感建模。虽然 Reverse-JST 较 JST 有一定的优化,但是当情感先验信息被编码时,其在文本情感分析上的表现始终比 JST 的表现差。Dermouche 等人^[9]提出了“自底向上”的 JTS 模型,这一点与 Reverse-JST 类似。JTS 模型无须后处理过程来匹配来自不同情感极性的主题,同时能够动态地修改先验参数,避免了使用固定先验参数而出现性能下降的缺陷。黄俊衡等人^[10]提出一种新的情感主题混合模型,即时间用户情感模型 (Time-User Sentiment Latent Dirichlet Allocation, TUS-LDA)。其本质可以看作是将同一时间下或同一用户发出的帖子聚合成一个伪的长文档,丰富上下文信息,避免了短文本数据稀疏的问题,从而可以学习到高质量的情感相关的主题。虽然 TUS-LDA 模型结合时间信息和用户信息来进行建模,但是该模型需要分别增加同一时间段以及同一用户下的主题分布,且在模型中引入了过多的参数加大了模型的复杂度。黄发良等人^[11]将用户性格特征和情感表情符号与 JST 模型相结合建立了 TSMMF (Topic Sentiment Model Based on Multi-feature Fusion),利用正负情感微博比例来量化微博用户性格情绪特征。这种性格情绪参数的建模方式具有良好的扩展性,但是比较抽象简单,可以纳入更多的因子来全面刻画用户性格情绪特性。许银洁等人^[12]将用户统计特征和行为事件特征加入 JST 模型中,

提出了 JUST (Joint-User Sentiment/Topic) 模型, 在汽车评论数据集上进行了实验, 效果优于 TSMMF 模型。但是该模型只是将汽车评论数据集中的用户年龄和性别等人口统计特征量化为用户特征, 且并没有在真实的微博数据集下对比情感分类的效果, 因而该模型在微博情感分析领域的有效性还有待考证。

综上所述, 主题情感模型虽然在情感分析领域得到了一定的应用, 但是仍存在不足。本文提出一种基于时间的用户性格情绪特征建模方法, 将用户性格情绪特征纳入主题情感模型中, 同时在模型中融入了表情符号元素, 进而建立起了 UC-JST 模型。

2 UC-JST 主题情感模型

UC-JST 主题情感模型是在 JST 模型中融入了用户性格情绪特征及表情符号信息, 通过新增用户性格情绪参数在模型迭代推导过程中对微博消息的情感进行调整, 引入表情符号信息以更准确地捕捉微博信息的隐含情感, 从而加强主题情感模型对微博情感的挖掘质量。

为此, 首先我们需要对用户性格情绪特征建模, 然后将建模好的性格情绪特征参数加入模型的迭代过程中。

2.1 性格情绪特征建模

对性格进行建模的方式并不多, 其中大多数都是基于心理学来实现的。现有的性格模型都是多维的, 例如, 著名的三因素模型、七因素模型、Big Five 模型^[13]、“性格—心情—情感”模型^[14], 以及基于性格、情感空间的 HMM 多层次情感计算模型^[15]等, 都从心理学的角度对性格进行了量化研究, 但是这些模型都是基于小规模的用户数据调查进行的分析, 基于类似微博文本这种大规模数据的研究成果很少。对个体而言, 性格影响着他的积极与消极情绪的变化范围和变化率, 也就是说, 一个时间步长内, 情感变化多快, 以及变化到何种程度, 是性格的函数, 并且“情感一致性”理论^[16]表明: 考察某条微博的情感时, 其相邻时间内的微博也具有一定的参考价值, 但是目前的这些模型却忽略了时间这个因素。

对性格的各项基本特征进行分析后, 鉴于本文讨论的是微博文本的情感分析方法, 对性格进行全方位建模是不切实际的, 受文献^[17]的启发, 选取性格各项特征中的情绪特征来基于时间进行建模。

在主题情感模型中, 本文对性格情绪参数定义如下: 对于微博博主 u 发表的所有微博, 对于情感极性 s 来说, 性格情绪参数 $\lambda_{u,s}$, 定义如式(1)所示。

$$\lambda_{u,s} = \frac{\sum_{m_i \in C_{u,s}} (\text{time} - |t_{m_0} - t_{m_i}|)}{\sum_{m_i \in C_u} (\text{time} - |t_{m_0} - t_{m_i}|)} \quad (1)$$

其中, C_u 表示在时间阈值 time 内该用户发表的全部微博集合, $C_{u,s}$ 表示 C_u 中具有情感标签 s 的微博集合, m_0 表示当前待考察的微博, m_i 表示 C_u 中除 m_0 以外的任意一条微博, t_{m_0} 表示微博 m_0 所发表的时间, t_{m_i} 表示微博 m_i 的发表时间。 $\lambda_{u,s}$ 的定义遵循规则“考量微博 m_0 时, 以 m_0 为时间轴中心, 微博 m_i 的权值随着距离 m_0 发表时间间隔变大呈现梯度递减”, 模型中不仅考虑了发表时间这个因素, 且区分了发表时间远近对微博情感的影响程度。

本文将构建好的性格情绪参数 λ 加入到模型迭代过程中。由于在模型参数推理时, 微博主题情感相关分布是动态变化的, 因此性格情绪参数 λ 是自适应调节的。

2.2 表情符号

微博短文本受到字数限制, 有时并不能完整表达用户的情感和态度, 而表情符号不但自身具有情感倾向, 也能影响微博整体的情感倾向性。除此之外, 表情符号在不同的使用情况下有不同的引申义, 能够传达出丰富的语义信息, 这是文字表达所不能企及的。近年来也有不少学者开始研究如何充分利用表情符号来提升微博情感识别性能, 吴晨茜等人^[18]提出了一种融合表情符号和情感词的自动标注方法来减轻人工标注的工作量, 同时将表情符号向量化, 并训练情感分类器进行微博情感分析, 提高了情感分类的准确率。何炎祥等人^[19]将基于表情符号的情感空间映射与深度学习模型 MCNN 结合, 有效增强了 MCNN 捕捉情感语义的能力, 在 NLPCC 的微博情感分类数据集上取得了非常好的分类效果。

受上述研究启发, 为了更加客观、准确地判断微博的情感倾向, 本文在 JST 模型基础上引入了表情符号变量, 使得微博消息组成元素在单词 w 基础上增加了表情符号 e 。

2.3 模型描述

通过在 JST 模型中引入用户性格情绪特征和表情符号, 本文提出了一个改进的主题模型 UC-JST。

UC-JST 的概率图模型如图 1 所示,相关符号及说明如表 1 所示。

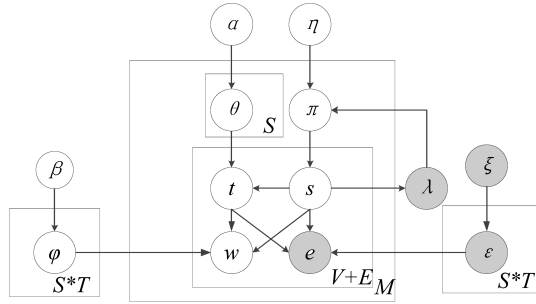


图 1 UC-JST 模型

表 1 UC-JST 模型符号含义

符号	含义	符号	含义
M	微博消息集的微博数	α	Dirichlet 分布 θ 的先验参数
V	微博消息集的词典大小	β	Dirichle 分布 φ 的先验参数
E	微博消息集的情感标签库大小	η	Dirichlet 分布 π 的先验参数
T	微博消息集隐含的主题数	ξ	Dirichlet 分布 ϵ 的先验参数
S	微博情感类别数	λ	Dirichlet 分布 π 的性格参数
U	微博消息集的用户总计数		
π	微博-情感分布	s	情感
θ	微博、情感—主题分布	t	主题
φ	情感、主题—词分布	w	单词
ϵ	情感、主题—情感符号分布	e	情感符号

UC-JST 模型的基本思想如下：假设微博消息集中有 M 条微博,表示为 $C = \{m_1, m_2, \dots, m_M\}$; 微博 m_i 是由 V_i 个单词和 E_i 个表情符号组成的序列,记为: $m_i = \{w_1, w_2, \dots, w_{V_i}; e_1, e_2, \dots, e_{E_i}\}$ 。微博语料库中去重后的词汇表中的词汇数量为 V , 则微博中每个词对应 V 中的一个索引项; 微博消息集中隐含的主题个数为 T , 情感类别个数为 S , 情感标签个数为 E , 微博用户总数为 U 。在 UC-JST 模型中, 情感分布 π 通过 η 为超参数的 Dirichlet 先验分布产生; 同理, $\theta, \varphi, \epsilon$ 分别服从 Dirichlet 分布 $\text{dir}(\alpha), \text{dir}(\beta), \text{dir}(\xi)$ 。在 UC-JST 模型中, 相比于 JST 模型主要进行了两个方面的改进: ①引入表情符号 e 以更准确地捕捉微博消息的隐含情感; ②新增性格情绪参数 λ 以在模型迭代推导微博主题情感

的过程中对微博消息的情感进行调整。

UC-JST 模型定义的微博生成过程是概率抽样的过程, 其形式化定义如下:

- 对于每篇微博 m , 生成微博的情感分布 $\pi_m \sim \text{dir}(\beta)$ 。对于每种情感标签 s , 生成微博的情感-主题分布 $\theta_{m,s} \sim \text{dir}(\alpha)$ 。
- 对于每个主题情感对 (t, s) , 生成主题情感对的单词分布 $\varphi_{t,s} \sim \text{dir}(\beta)$ 。生成主题情感对的情感符号分布 $\epsilon_{t,s} \sim \text{dir}(\xi)$ 。
- 对微博中的每个词语 w , 选择一个情感标签 $s \sim \text{multinomial}(\pi_m)$, 然后选择一个主题 $t \sim \text{multinomial}(\theta_{m,s})$, 最后通过 $w_{t,s} \sim \text{multinomial}(\varphi_{t,s})$ 获取词语; 而对微博中的每个情感符号 e , 最后通过 $e_{t,s} \sim \text{multinomial}(\epsilon_{t,s})$ 获取表情符号。其中, $\text{multinomial}()$ 表示多项式分布。

2.4 UC-JST 模型推导

UC-JST 模型推断的关键问题在于求解每一篇微博 m 的情感分布 π_m 、每一篇微博 m 下的每一个情感标签 s 的主题分布 $\theta_{m,s}$, 以及每一个情感标签和主题标签下的词语分布 $\varphi_{t,s}$ 或者情感符号分布 $\epsilon_{t,s}$ 。本文采用吉布斯采样法^[20]来估计 UC-JST 模型参数, 吉布斯采样是一种快速高效的 MCMC (Markov Chain Monte Carlo) 采样方法, 可以通过不断地迭代采样对概率模型的参数进行推导。为了估计以上四个分布, 根据吉布斯采样的方法, 需要先对条件概率分布进行估计, 如式(2)所示。

$$P(t_i = t, s_i = s \mid t_{-i}, s_{-i}, w, e, \alpha, \beta, \eta, \xi, \lambda) \quad (2)$$

其中, t_i 表示微博 m 中第 i 个位置上的元素(词语或者表情符号)的主题, t_{-i} 表示微博 m 中除了第 i 个位置上的元素以外的其他元素的主题; s_i 表示微博 m 中第 i 个位置上的元素的情感极性, s_{-i} 表示微博 m 中除了第 i 个位置上的元素以外的其他元素的情感极性。

要估计上述条件概率公式, 必须先计算联合概率公式 $P(w, e, t, s)$, 根据条件概率公式基本定理可知:

$$P(w, e, t, s) = P(w \mid t, s)P(e \mid t, s)P(t \mid s)P(s) \quad (3)$$

要想求解 $P(w \mid t, s), P(e \mid t, s), P(t \mid s), P(s)$ 各个因子的值, 必须通过对 φ 分布、 ϵ 分布、 θ 分布和 π 分布分别进行积分运算。具体如式(4)~式(7)所示。

$$P(w | t, s) = \int P(w | t, s, \varphi) P(\varphi | \beta) d\varphi$$

$$= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^{S \times T} \prod_{s=1}^S \prod_{t=1}^T \frac{\prod_{w=1}^V \Gamma(N_{w,t,s} + \beta)}{\Gamma(N_{t,s} + V\beta)} \quad (4)$$

式(4)中, $N_{w,t,s}$ 表示词语 w 在情感标签 s 和主题标签 t 下出现的次数, $N_{t,s}$ 表示所有词语在情感标签 s 和主题标签 t 下出现的总数, $\Gamma(\cdot)$ 表示伽马函数。

$$P(e | t, s) = \int P(e | t, s, \epsilon) P(\epsilon | \xi) d\epsilon$$

$$= \left(\frac{\Gamma(E\xi)}{\Gamma(\xi)^E} \right)^{S \times T} \prod_{s=1}^S \prod_{t=1}^T \frac{\prod_{e=1}^E \Gamma(M_{e,t,s} + \xi)}{\Gamma(M_{t,s} + E\xi)} \quad (5)$$

式(5)中, $M_{e,t,s}$ 表示表情符号 e 在情感标签 s 和主题标签 t 下出现的次数, $M_{t,s}$ 表示所有表情符号在情感标签 s 和主题标签 t 下出现的总数。

$$P(t | s) = \int P(t | s, \theta) P(\theta | \alpha) d\theta$$

$$= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \right)^{S \times M} \prod_{m=1}^M \prod_{s=1}^S \frac{\prod_{t=1}^T \Gamma(N_{m,t,s} + \alpha)}{\Gamma(N_{m,s} + T\alpha)} \quad (6)$$

式(6)中, $N_{m,t,s}$ 表示微博消息 m 中的当前元素同时出现在情感标签 s 和主题标签 t 下的次数, $N_{m,s}$ 表示微博消息 m 中的所有元素出现在情感标签 s 下的次数。

$$P(s) = \int P(s | \pi) P(\pi | \eta) d\pi$$

$$= \left(\frac{\Gamma(S(\eta + \lambda))}{[\Gamma(\eta + \lambda)]^S} \right)^M \prod_{m=1}^M \frac{\prod_{s=1}^S \Gamma(N_{m,s} + \eta + \lambda_s)}{\Gamma(N_m + S(\eta + \lambda_s))} \quad (7)$$

式(7)中, N_m 表示微博消息 m 中元素(词语或者表情符号)的总个数, λ_s 表示用户的性格情绪参数, 通过嵌入在微博的情感分布中, 会在吉布斯采样过程中对微博消息的情感进行调整。

根据式(4)~式(7), 可以计算联合概率 $P(w, e, t, s)$, 从而进一步得到吉布斯采样所需的条件概率, 如式(8)所示。

$$P(t_i = t, s_i = s | t_{-i}, s_{-i}, w, e, \alpha, \beta, \eta, \xi, \lambda)$$

$$\propto \frac{(N_{m,t,s})_{-i} + \alpha}{(N_{m,s})_{-i} + T\alpha} \times \frac{(N_{w,t,s})_{-i} + \beta}{(N_{t,s})_{-i} + V\beta}$$

$$\times \frac{(M_{e,t,s})_{-i} + \xi}{(M_{t,s})_{-i} + E\xi} \times \frac{(N_{m,s})_{-i} + \eta + \lambda_s}{(N_m)_{-i} + S(\eta + \lambda_s)} \quad (8)$$

式(8)中, “ $-i$ ”表示排除当前位置的元素, 例如, $(N_{m,t,s})_{-i}$ 表示在微博 m 中, 除了第 i 个位置上的元素, 其他元素在情感标签 s 和主题标签 t 下出现的次数。其他次数项同理解释。

通过式(8)计算出微博中每一个词语或者情感符号在不同的主题和情感下的概率分布, 并采样出情感和主题标签。经过迭代多次后, 分布达到细致平稳, 即收敛状态, 根据采样的样本, 可以估计出模型的参数。这些参数通过狄利克雷分布的期望值得到, 其中词语的分布、情感符号的分布、微博的主题分布和微博的情感分布分别如式(9)~式(12)所示。

$$\varphi = \frac{N_{w,t,s} + \beta}{N_{t,s} + V\beta} \quad (9)$$

$$\epsilon = \frac{M_{e,t,s} + \xi}{M_{t,s} + E\xi} \quad (10)$$

$$\theta = \frac{N_{m,t,s} + \alpha}{N_{m,s} + T\alpha} \quad (11)$$

$$\pi = \frac{N_{m,s} + \eta + \lambda_s}{N_m + S(\eta + \lambda_s)} \quad (12)$$

利用微博的情感分布 π , 可以得到微博在情感标签 s 下的概率, 从而获取微博的情感极性。

2.5 UC-JST 模型先验

主题情感模型中, 通常会引入先验知识来给模型进行初始化, 先验知识利用得好, 会大大提升模型的效率。本文在 UC-JST 模型中引入两种先验知识, 即词语情感先验和表情符号先验。这里需要说明的是, 词语情感先验和表情符号先验是为了给主题情感模型更好的初始化, 有效提高无监督学习算法的性能, 其不同于监督学习中的标注数据, 因此需要区分开。

(1) 词语情感先验

微博是一个跨领域社交平台, 其中的文本涉及到的领域知识非常广泛, 因此不能选取单一的情感词典来给词语赋值。本文选取应用较为广泛的三个情感词典“大连理工大学中文情感词汇本体库”“台湾大学中文情感极性词典(NTUSD)”“知网情感词典(HOWNET)”, 根据以上三个情感词典对微博中的词语赋予情感极性。

(2) 表情符号先验

UC-JST 模型中需要使用到微博表情符号, 目前(2019 年)网页版微博支持的默认表情符号有 84 个, 手机版微博支持的默认表情符号有 160 个, 其中

只有少数表情符号经常被使用。考虑到表情符号种类增加的同时会带来模型的算法时间复杂度的增加,数量级为 $O(M \times (V + E))$, 结合表情符号的使用情况, 本文选取 28 个常用表情符号, 构成表情符号库。这 28 个表情符具有很强的代表性, 在我们所有的表情符号总计数中占 95% 以上。由于受篇幅影响, 这里只分别列举三组积极和消极的表情符号, 如表 2 所示。

表 2 微博表情符号库

符号	文字表示	符号	文字表示
	[可爱]		[抓狂]
	[鼓掌]		[失望]
	[哈哈]		[怒骂]
...		...	

3 实验与分析

3.1 数据集

目前虽然有许多学者对中文微博情感分析这项工作进行研究, 但是针对基于主题模型的中文微博情感分类的研究却非常少, 本文需要用到微博文本的发表时间、用户 id、表情符号等信息, 现有的数据集不满足要求, 因此从新浪微博网站上按照用户活跃度爬取活跃度较高的用户 (在一个采集周期 (10d) 内所发微博数目大于 3 的用户) 微博数据, 筛选后进行实验数据集构造。根据不同的采集时间, 本文构造了三个数据集: 时间段 (2018/10/1 ~ 2018/10/10) 内的数据集 1、时间段 (2019/1/1 ~ 2019/1/10) 内的数据集 2 以及时间段 (2019/6/16 ~ 2019/6/25) 内的数据集 3, 如表 3 所示。

表 3 实验数据集

数据集	训练集 (积极/消极)	测试集 (积极/消极)
数据集 1	6 500 条/6 325 条	4 450 条/4 200 条
数据集 2	6 300 条/6 210 条	4 320 条/4 100 条
数据集 3	6 610 条/6 550 条	4 640 条/4 450 条

微博用户可自由发表自己的观点和见解, 其表达基本不受限制, 语法格式有时并不符合正常的规范。鉴于此, 本文需要对爬取的微博数据做以下预处理操作: ①去除文本中的英文和数字; ②分词时保留表情

符号; ③过滤停用词。最终每一条微博数据包含以下基本信息: 微博正文、发表时间、用户 id。若微博中存在表情符号, 那么还应包含表情符号。

由于爬取的微博数据没有已知的情感极性, 所以本文通过人工标注的方式对数据集进行情感标注。在情感标注过程中, 3 位情感标注者分别对数据集进行独立标注, 最终使用 Kappa 一致性检验对标注结果进行了检验, 检验结果表明一致性达到了 90% 以上。对于标注结果不统一的数据, 使用投票法进行判断。

3.2 评价指标

在文本情感分类领域, 评价二分类模型性能的常用指标为准确率 Precision (P)、召回率 Recall (R) 以及 F_1 值。 P 指关注类的正确分类文本数占测试集中分到该类的文本数量的比例, R 指模型对关注类的正确分类的文本数占测试集该类实际文本数的比例。而 F_1 值是一个综合度量指标, 是准确率和召回率的调和平均值, F_1 值越高, 模型的分类效果越好。本文最终选取 F_1 值来评价主题情感模型进行微博情感分类的效果。

3.3 实验过程

3.3.1 先验参数设置

UC-JST 模型中的先验参数参考 JST 模型^[2]的做法来设置, 对称超参数的设定为 $\alpha = 50/\text{主题数}$, $\beta = 0.01$; 不对称超参数的设定中, 对于积极情感, $\eta = 0.01$, 对于消极情感, $\eta = 5$ 。另外, 先验参数 ξ 按照 TSMMF 模型^[11]的做法设置, $\xi = 0.01$ 。对比模型中的超参数以其作者所提供的为准。设置参数 α 时, 主题数目直接影响到 α 的初始值, 为此本文每组实验将在不同的主题数目下进行迭代, 并取在最优主题数目下获取的结果。

3.3.2 对比实验及结果分析

为了定量地分析 UC-JST 模型在微博情感分类任务中的性能, 本节首先分析引入性格情绪参数对情感分类的影响, 然后再探讨表情符号的占比对情感分类的影响, 最后再选取四个典型的基于 JST 模型改进的无监督情感分类算法 JST、TUS-LDA、JUST、TSMMF 与 UC-JST 模型进行对比实验。

(1) 性格情绪参数对情感分类的影响

用户性格情绪参数的引入, 是 UC-JST 模型相对于 JST 模型来说非常重要的改进, 本组实验去除了微博中的所有表情符号, 考察性格情绪参数对情

感分类的影响,具体分为两个方面:①不同的时间阈值对情感分类的影响,这里的时间阈值指的是 2.1 节中的时间阈值 time;②性格情绪参数的引入对情感分类的影响。

① 针对不同的时间阈值对情感分类的影响,设置了 12 个时间阈值,最小的为 6h,最大的为 72h,在最优主题数目条件下,分别在三组数据集上各进行 10 轮实验,得出结果如图 2 所示。

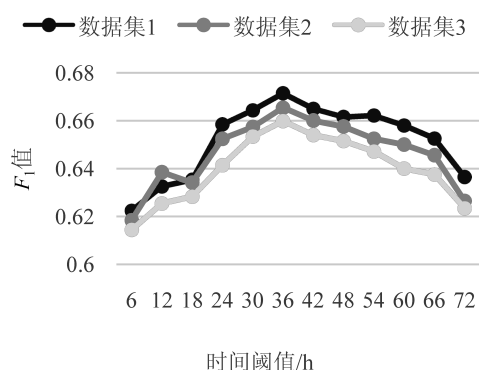


图 2 不同时间阈值对情感分类的影响

从图 2 中可以看出,随着时间阈值从 6h 至 36h 逐渐增大, F_1 值逐渐增大,但是当时间阈值大于 36h 后, F_1 值又呈现下降趋势。分析可知,当时间阈值较小时,阈值内的微博数目较少,对待考量的微博的情感影响较小,因此情感分类效果一般;当时间阈值太大时,阈值内的微博数目虽然足够多,但是这些微博与待考量微博的情感联系不够紧密,会造成反向的情感误差,因此恰当的时间阈值是模型中非常重要的参数。根据实验结果,本文选取 36h 作为 UC-JST 模型的时间阈值。

② 针对性格情绪参数的引入,分别设置了“未引入性格情绪参数”和“引入性格情绪参数”两组实验,结合上文的实验结果,选取 36h 作为本实验的时间阈值,在最优主题数目条件下,在数据集 1 上进行 10 轮实验,得到的实验结果如图 3 所示。

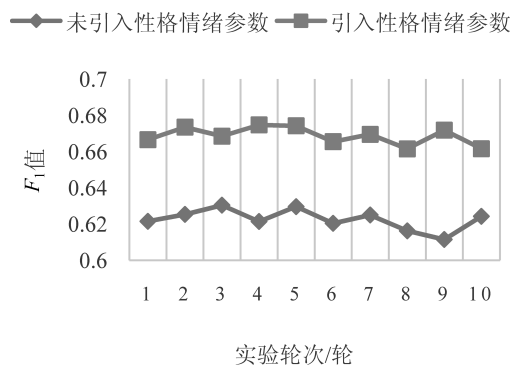
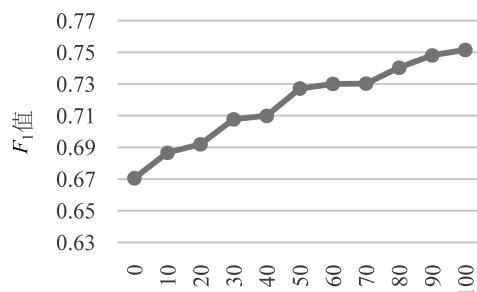


图 3 数据集 1 上用户性格情绪参数对情感分类的影响

由图中可以看出,两组对比实验中,引入用户性格情绪参数的 F_1 值明显高于未引入用户性格情绪参数的 F_1 值,情感分类的效果有所提升。经过计算,未引入性格情绪参数的平均 F_1 值约为 0.62,引入性格情绪参数的平均 F_1 值约为 0.67,足以说明性格情绪参数的引入对中文微博情感分类有一定的帮助作用,验证了引入用户性格情绪特征的有效性。

(2) 表情符号对情感分类的影响

表情符号也是 UC-JST 模型中重要的微博特征之一,本组实验考察微博中表情符号对情感分类结果的影响。由于微博文本中有的包含表情符号,有的不包含,因此本实验考察语料库中包含表情符号的微博占有所有微博的条数比重对情感分类的影响。实验结果如图 4 所示。



带表情符号的微博数目占比/%

图 4 数据集 1 上表情符号对情感分类的影响

从图 4 中可以看出,极端情况下,当语料库中的微博全部不包含表情符号时, F_1 值约为 0.67,情感分类效果较差;随着带表情符号的微博占比增加,情感分类的效果也越来越好,二者基本成正比关系;当所有微博都包含表情符号时, F_1 值约达到了 0.75,这在无监督学习的情感分类方法中,是比较出色的效果。综合实验结果说明,表情符号的引入对情感分析结果有着良好的促进作用,究其原因,表情符号在模型先验中就赋予了特定的情感极性,而不是随机值,对于模型的训练过程有着较为明显的帮助作用。

(3) 情感分类结果对比分析

通过以上几组实验,将 UC-JST 模型中的重要参数确定下来后,本文选取 JST、TUS-LDA、JUST、TSMMF 这四个具有代表性的无监督情感学习模型与 UC-JST 模型进行对比,在最优主题数目及时间阈值为 36h 的前提下,分别在三组不同的数据集上各进行了 10 组实验,实验结果如表 4~表 6 所示。

表 4 数据集 1 上情感分类结果对比分析

组号	JST	TUS-LDA	JUST	TSMMF	UC-JST
1	0.617	0.658	0.701	0.709	0.754
2	0.618	0.653	0.709	0.704	0.752
3	0.609	0.668	0.702	0.711	0.751
4	0.605	0.641	0.697	0.708	0.744
5	0.608	0.665	0.699	0.718	0.753
6	0.621	0.655	0.704	0.715	0.750
7	0.619	0.642	0.693	0.712	0.745
8	0.605	0.669	0.704	0.699	0.758
9	0.607	0.657	0.710	0.703	0.749
10	0.612	0.663	0.698	0.704	0.752
Avg(F_1)	0.612	0.657	0.702	0.708	0.751

表 5 数据集 2 上情感分类结果对比分析

组号	JST	TUS-LDA	JUST	TSMMF	UC-JST
1	0.603	0.654	0.681	0.703	0.749
2	0.594	0.652	0.683	0.695	0.743
3	0.598	0.656	0.678	0.702	0.739
4	0.601	0.642	0.677	0.698	0.756
5	0.602	0.645	0.691	0.707	0.741
6	0.606	0.647	0.688	0.699	0.744
7	0.608	0.643	0.692	0.691	0.738
8	0.604	0.651	0.682	0.693	0.747
9	0.595	0.669	0.689	0.694	0.754
10	0.597	0.658	0.696	0.701	0.745
Avg(F_1)	0.601	0.652	0.686	0.698	0.746

表 6 数据集 3 上情感分类结果对比分析

组号	JST	TUS-LDA	JUST	TSMMF	UC-JST
1	0.606	0.658	0.681	0.682	0.740
2	0.602	0.661	0.685	0.685	0.736
3	0.601	0.663	0.691	0.694	0.741
4	0.612	0.672	0.671	0.677	0.739
5	0.599	0.667	0.674	0.692	0.743
6	0.597	0.671	0.682	0.678	0.738
7	0.613	0.666	0.681	0.683	0.727
8	0.605	0.678	0.686	0.688	0.734
9	0.596	0.675	0.689	0.696	0.742
10	0.607	0.667	0.684	0.684	0.733
Avg(F_1)	0.604	0.668	0.682	0.686	0.737

经过在三组数据集上的实验对比可知,UC-JST 模型在数据集 1、数据集 2、数据集 3 上的平均 F_1 值分别达到了 0.751、0.746 和 0.737,造成这种细微差距的原因在于每个数据集中的表情符号占比不同。通过进一步对比发现,无论是平均 F_1 值还是最大 F_1 值,UC-JST 模型均较其他四个模型有明显的优势,而且较 TSMMF 模型的平均 F_1 值提升了接近 0.05。这说明表情符号和用户性格情绪参数的引入对于情感分类有着非常重要的作用,体现了基于时间的性格情绪建模方式的可靠性,也验证了 UC-JST 模型在情感分类领域的可行性和有效性。

4 结论

本文主要针对 JST 模型在中文微博情感分类领域建模时微博特征缺失的问题,在该模型中引入了表情符号和用户性格情绪特征,并提出了一种基于时间的用户性格情绪特征建模方法,进而提出了 UC-JST 模型。然后通过实验确定了 UC-JST 模型的时间阈值这个重要参数,定量、定性地分析了引入的特征对微博情感分类效果的影响,最后通过对比实验验证了该模型的可行性及有效性。

虽然 UC-JST 模型作为无监督学习方法在中文微博情感分类任务上表现不错,但是仍然存在一些不足,本文的情感极性只包含积极和消极两种,在 UC-JST 模型基础上进行中文微博的细粒度情感分类将会是下一步的工作。

参考文献

[1] 黄发良,于戈,张继连,等. 基于社交关系的微博主题情感挖掘[J].软件学报. 2017, 28(03): 694-707.

[2] Lin C, He Y. Joint sentiment/topic model for sentiment analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009: 375-384.

[3] Peterson C. A primer in positive psychology[M]. Oxford: Oxford University Press, 2006: 182-208.

[4] 杜慧,陈云芳,张伟. 主题模型中的参数估计方法综述[J]. 计算机科学, 2017, 44(S1):29-32,47.

[5] 蔡永明,长青.共词网络 LDA 模型的中文短文本主题分析[J].情报学报,2018,37(03):305-317.

[6] Nguyen D Q, Billingsley R, Du L, et al. Improving topic models with latent feature word representations [J]. Transactions of the Association for Computational Linguistics, 2015, 3(1):299-313.

- [7] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26(1): 3111-3119.
- [8] Lin C, He Y, Everson R, et al. Weakly supervised joint sentiment-topic detection from text[J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24(6): 1134-1145.
- [9] Dermouche M, Kouas L, Velcin J, et al. A joint model for topic-sentiment modeling from text[C]//Proceedings of the 30th Annual ACM Symposium on Applied Computing, 2015: 819-824.
- [10] 黄俊衡. 基于改进主题模型的微博短文本情感分析的研究[D]. 南京: 东南大学硕士学位论文, 2017.
- [11] 黄发良, 冯时, 王大玲, 等. 基于多特征融合的微博主题情感挖掘[J]. 计算机学报, 2017, 40(04): 872-888.
- [12] 许银洁, 孙春华, 刘业政. 考虑用户特征的主题情感联合模型[J]. 计算机应用, 2018, 38(05): 1261-1266.
- [13] 袁婷婷, 杨文忠, 仲丽君, 等. 一种基于性格的微博情感分析模型 PLSTM[J/OL]. 计算机应用研究, 2019, 37(02): 1-6.
- [14] 李海芳, 何海鹏, 陈俊杰. 性格、心情和情感的多层情感建模方法[J]. 计算机辅助设计与图形学学报, 2011, 23(04): 725-730.
- [15] Kshirsagar S, Magnenat-Thalman N. A multilayer personality model[C]//Proceedings of the 2nd International Symposium on Smart Graphics, 2002: 107-115.
- [16] Abelson R P. Whatever became of consistency theory? [J]. Personality & Social Psychology Bulletin, 1983, 9(1): 37-64.
- [17] 赵蓉英, 张扬. 基于时空维度的国内外情感分析研究演化分析[J]. 情报科学, 2018, 36(10): 171-177.
- [18] 吴晨茜, 陈锻生. 表情符向量化算法[J]. 华侨大学学报(自然科学版), 2019, 40(03): 399-404.
- [19] 何炎祥, 孙松涛, 牛菲菲, 等. 用于微博情感分析的一种情感语义增强的深度学习模型[J]. 计算机学报, 2017, 40(04): 773-790.
- [20] He L, Jia Y, Han W, et al. Mining user interest in microblogs with a user-topic model[J]. China Communications, 2014, 11(8): 131-144.



李玉强(1977—), 博士, 副教授, 主要研究领域为机器学习与大数据分析。

E-mail: liyuqiang@whut.edu.cn



孙念(1994—), 硕士, 主要研究领域为自然语言处理。

E-mail: 853128958@qq.com



黄瑜(1995—), 通信作者, 硕士研究生, 主要研究领域为自然语言处理。

E-mail: 1391460498@qq.com