北京大学学报(自然科学版) 第 54 卷 第 4 期 2018 年 7 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 4 (July 2018)
doi: 10.13209/j.0479-8023.2018.011

游客微博主题情感分析方法比较研究

刘思叶 田原† 冯雨宁 庄育龙

北京大学遥感与地理信息系统研究所, 北京 100871; † 通信作者, E-mail: tianyuanpku@pku.edu.cn

摘要 针对饮食、娱乐、购物、景观、交通和住宿6个旅游主题, 基于机器学习方法, 开展游客微博主题情感分析方法比较研究。以人工标注的53140条赴日游客微博为数据基础, 应用两种机器学习模型开展建模实验, 并分析不同特征对建模效果的影响。实验结果显示, 两种模型的建模效果良好, 适用于游客微博主题情感分析, 其中最大熵模型效果略优于支持向量机。研究还表明, 在词特征的基础上引入表情符号和主题词进行特征扩展, 可以提高模型的建模效果。

关键词 主题情感分析; 游客微博; 最大熵模型; 支持向量机

中图分类号 F590

Comparison of Tourist Thematic Sentiment Analysis Methods Based on Weibo Data

LIU Siye, TIAN Yuan†, FENG Yuning, ZHUANG Yulong

Institute of Remote Sensing and Geographical Information System, Peking University, Beijing 100871;
† Corresponding author, E-mail: tianyuanpku@pku.edu.cn

Abstract Six tourism themes, diet, entertainment, shopping, view, transportation, and accommodation, are selected for thematic sentiment analysis. 53140 Weibo items published by Chinese tourists in Japan are collected and manually labeled as the case study dataset. Maximum Entropy model and Support Vector Machine are adopted. The training results are both fairly good, where the resulting Maximum Entropy model prevails slightly. It can be concluded that machine learning models are reasonably feasible in tourist thematic sentiment analysis. Moreover, the experiment also shows that the models can be improved by introducing emoticon icons and thematic words as supplements to traditional word features.

Key words thematic sentiment analysis; Weibo of tourists; Maximum Entropy; Support Vector Machine (SVM)

随着互联网的快速发展与普及, 微博作为社交自媒体, 已经成为我国游客分享旅途见闻与感受的重要平台。游客微博通常具有一定的情感色彩[1], 通过分析其中包含的情感信息, 可以了解游客对旅游目的地的满意程度, 对旅游管理与决策有重要意义[2]。

情感分析是对文本信息进行分析和挖掘, 可以识别出文本所表达的积极或消极情感[3–4], 是常用的文本分析方法之一。主题情感分析是情感分析的子领域, 即识别出文本信息针对特定主题的情感倾向[5]。主题情感分析能够区分同一文本信息针对不同主题的情感倾向, 因而可以挖掘出更丰富的语义信息。

目前有关游客微博主题情感分析的研究较少, 尚无完善的情感标注数据集和统一的分析方法。已有的研究主要采用基于情感词典的计算方法, 机器学习方法尚未得到充分应用[2,5]。基于情感词典的计算方法, 依赖人工设定的规则和情感词典, 仅适用于部分样本, 当文本结构较复杂或包含未收录的情感词时, 常常无法正确处理[3–4]。机器学习方法

国家重点研发计划(2018YFB0505500, 2018YFB0505504)和测绘遥感信息工程国家重点实验室开放研究基金((16)重02)资助
收稿日期: 2017–05–05; 修回日期: 2017–06–07; 网络出版日期: 2018–04–18

可以从训练数据中发现规律并构建分类器，进而应用于测试数据的分类[6]。这类方法不需要人工定义规则，概括能力较强，能够应用于大规模的数据处理，同时可以对分析结果进行定量评价。

本文以我国赴日游客微博为研究实例，通过API 调用和网络爬虫采集赴日游客微博的文本数据，对数据进行人工标注，获得标注数据集；然后基于向量空间模型，对标注数据集中的文本进行词特征表示，将变长的文本信息转为固定长度的特征向量，并引入微博表情和主题词进行特征扩展；最后采用最大熵模型(Maximum Entropy, MaxEnt)和支持向量机(Support Vector Machine, SVM)进行建模与效果分析，并对比不同特征对建模效果的影响。

1 数据获取与标注

考虑到境外游期间游客的身份比较容易确认，本文选取中国游客赴日旅游期间的微博作为研究数据。研究表明，在国人出国游目的地中，日本排名第三，属于热门旅游目的地之一，每年前往日本的游客数在 400 万人次以上[7]。以日本作为研究区域，具有较好的代表性。

1.1 数据采集与筛选

数据采集与筛选主要由 3 个步骤组成：原始数据采集、游客身份识别和微博信息筛选。

原始数据采集 本文使用新浪微博的 placeAPI①(能够获取位于某个点位周边一段时间内的微博签到信息)，以 2000 m 为半径，对整个日本地区全覆盖采集，提取 2015 年 12 月 1 日至 2016 年 11 月 30 日在日本签到的用户 ID。以采集到的用户 ID 为种子，通过网络爬虫获取用户发布的全部微博信息，包括时间、位置和文本。

游客身份识别 本文通过用户自描述信息和时空轨迹信息依次进行筛选，将原始数据中的中国游客与其他用户(如他国游客、本地居民和海外留学生等)区分开来，识别出中国游客用户，并将这些用户于 2015 年 12 月 1 日至 2016 年 11 月 30 日在日本发表的微博构成原始微博数据集。具体筛选规则如下。

1) 用户自描述信息。用户自描述信息主要包含用户来源地和用户自我简介。首先筛除所有未注明来源地或来源地为非中国的用户；进而采用关键字匹配的方法，筛除简介中包含租房、淘宝、微信、留学等关键字的用户(这些账户大部分是中介或商家所持有)。

2) 时空轨迹信息。研究认为，通常一个游客在旅游地的滞留时间不超过30天[8–9]。本研究中，若用户微博显示该用户在日本停留超过一个月以上，则筛除该用户。此外，由于出海或 GPS 异常定位等原因，可能使得用户的位置信息不在日本本土范围内，针对这种情况，若其前后一天均有定位于日本的微博，则认为其定位也在日本。

微博信息筛选 原始微博数据集中存在无意义的微博数据，本文采用基于规则的方法筛除这些信息。具体规则如下。

1) 筛除包含淘宝、微信、租房等关键词的广告微博。

2) 由于中国常用字为简体汉字，筛除微博正文内容中包含日文或主要由英语以及繁体字构成的微博。

通过筛选，获得的数据集共包含 53140 条微博数据。

1.2 数据标注

徐海丽等[5]的研究表明，饮食、娱乐、购物、景观、交通和住宿是游客关心的主要方面，本文针对这 6 个主题开展研究。针对各个主题，我们将情感倾向分为三大类：积极、消极和中性[3]。积极指游客对该主题表达明显的正面情绪；消极指游客对该主题表达明显的负面情绪；中性指游客在微博中不涉及该主题内容，或未对该主题表达明显的积极或消极情感倾向，或对该主题表达出自相矛盾的情感倾向(如同时表达积极和消极情感)。

为了保证后续工作的可靠性，我们邀请两位近期曾赴日旅游的高校研究人员，对游客微博数据集的 53140 条微博数据进行人工标注，判别微博文本对 6 个主题的情感倾向。对两人标注结果不同的数据，通过讨论确定其主题情感倾向。表 1 展示不同主题标注结果的 Kappa 系数，平均值为 0.87。除景观主题外，Kappa 值均在 0.85 以上，这可能是由于游客对景观的情感表达较为间接或含蓄而导致。计算结果表明，两人的手工数据标注结果基本上一致，具有较好的可信度。表 2 展示不同主题的情感倾向分布，可以看出，游客对日本游的评价整体上较为

① https://api.weibo.com/2/place/nearby_timeline.json

表 1　人工标注结果 Kappa 系数
Table 1　Kappa coefficient of artificial labeling

主题	Kappa 值
饮食	0.91
娱乐	0.85
购物	0.87
景观	0.80
交通	0.89
住宿	0.90
平均	0.87

表 2　情感倾向分布
Table 2　Sentiment distribution

主题	积极	消极	中性
饮食	4406	604	48130
娱乐	1458	299	51383
购物	760	269	52111
景观	3224	270	49646
交通	322	337	52481
住宿	622	127	52391

正面，但对于交通方面，消极情绪的比例显著高于其他主题。

2　特征表示与扩展

2.1　基于向量空间的词特征表示

已有研究多采用向量空间模型，将微博表示为固定长度的特征向量[10]。$[f_1, f_2, f_3,..., f_i,..., f_m]$代表全部特征集合，对每个微博 d，将其表示为长度为 m 的向量$[n_1, n_2, n_3, ..., n_i, ..., n_m]$，其中 n_i 的取值范围为$[0, 1]$，代表特征 f_i 是否出现在微博 d 中[10]。

对每个主题，通过预处理与分词、特征提取和特征选择，获得微博文本针对该主题对应的特征向量，因此同一条微博文本对应 6 个特征向量，每个特征向量与一个主题相关。

1) 预处理与分词。微博文本存在较多非正式文本信息(如昵称、标签和统一资源定位符等)，这些内容无助于文本特征的表示和建模，需要去除。采用 jieba①分词工具，将处理后的文本分割为若干连续的词，每个词具有对应的词性。

2) 特征提取。我们以词特征作为基础特征，包括单词、连词、单词词性和连词词性。其中单词指文本中的单个词语，连词是指在文本中连续的两个单词，单词词性和连词词性分别是单词和连词所对应的词性信息[11]。通过特征提取，将分词后的微博文本转为由若干词特征组成的集合。

3) 特征选择。研究中包含的词特征数量往往是巨大的，若直接使用全部词特征，容易造成维度爆炸，降低模型的训练速度和效果[3]。卡方检验是常用的文本分类降维方法，其主要思想是词特征与某个类别之间符合卡方分布，卡方统计量越高，则该词特征与该类别之间的相关性越强，即通过该词特征能够更准确地判定文本是否属于该类别[12]。本文针对每个主题，在训练样本上计算每个词特征对积极、消极和中性 3 个情感类别的卡方统计量，采用这 3 个卡方统计量的平均值对全部词特征进行排序，选取前8000个词特征作为向量空间模型中的全部特征集合，将微博转为长度为8000的特征向量，并用于该主题下的情感分类。

2.2　表情符号特征扩展

中文微博网站提供大量表情符号，通过表情符号，用户能够传递和强调所表达的情感[13]。研究表明，引入表情符号特征能够提高情感分析模型的效果[13–14]。因此，我们将研究数据集中出现的403个表情符号分为积极、消极和中性三类，包含44个积极表情和 20 个消极表情。表 3 列举了一些具有代表性的积极、消极和中性表情。

2.3　主题词特征扩展

在主题情感分析中，某些词语与特定主题高度相关，这些词语往往只出现在与特定主题相关的文本中，称为主题词[15]。例如“好吃”一词，通常在人

表 3　微博表情示例
Table 3　Weibo emoticon examples

积极	消极	中性
		
		
		

① https://github.com/fxsjy/jieba

表 4 主题词示例

Table 4 Thematic word examples

主题	个数	示例
饮食	526	味道、拉面、牛肉
娱乐	120	温泉、迪士尼、好玩
购物	171	商场、售货员、买
景观	354	山景、竹林、樱花雨
交通	197	晕机、新干线、空姐
住宿	103	酒店、住宿、民居

们谈论饮食时使用，而在谈论交通或住宿等主题时较少使用。我们邀请两位标注人员，针对每个主题，对卡方检验中前2000个候选词进行人工识别和筛选，保留一致认为与该主题相关的词，得到对应的主题词集。表 4 展示不同主题的主题词示例。

3 建模与分析

本文选用最大熵模型和支持向量机开展建模训练，这两个模型都适用于文本分类、信息抽取和情感分析等任务[4]。最大熵模型是一种指数分类模型，在满足系统当前所有条件下，获取分布最平均的模型，即熵最大的模型[16]。支持向量机是基于核函数的分类模型，通过核函数将特征变换到线性可分的高维空间来完成分类[17]。

实验中，采用 scikit①软件包实现最大熵模型和支持向量机，其中支持向量机采用线性核函数。针对每个主题，我们分别训练了最大熵模型和支持向量机，共计 12 个模型实例，每个模型的输入是该主题下的特征向量，输出是微博针对该主题的情感倾向类别(即积极、消极或中性)。为了测试模型的泛化能力，我们对人工标注的 53140 条微博数据进行了随机划分，其中 60% 作为训练样本，用于特征选择和模型训练；40%作为测试样本，用于评价建模效果。

3.1 评价标准

不同模型间的建模效果采用 Macro-*F* 值进行对比，Macro-*F* 值越接近于 1，模型的效果越好[18]。计算公式如下，其中“+”表示积极情感，“−”表示消极情感。

$$准确率^{+(-)}=\frac{分类正确数^{+(-)}}{分类数^{+(-)}}, \tag{1}$$

$$召回率^{+(-)}=\frac{分类正确数^{+(-)}}{该类总数^{+(-)}}, \tag{2}$$

$$F值^{+(-)}=\frac{2\times准确率^{+(-)}\times召回率^{+(-)}}{准确率^{+(-)}+召回率^{+(-)}}, \tag{3}$$

$$\text{Macro-}F\text{值}=\frac{F值^{+}+F值^{-}}{2}。 \tag{4}$$

3.2 实验结果与分析

表 5 和 6 分别展示最大熵模型和支持向量机在不同特征组合上的建模结果。可以看出，单独引入表情特征或主题词特征均能提高建模效果，Macro-*F* 值比仅使用词特征均有所提高。在最大熵模型中，Macro-*F* 均值在仅使用词特征时为 0.329，加入表情特征后提高到 0.340，加入主题词特征后提高到 0.350；在支持向量机中，Macro-*F* 均值从 0.293 分别提升至 0.303 和 0.296。同时引入表情特征和主题词特征时，分类效果比之前又有全面的提升，最

表 5 最大熵模型 Macro-*F*值

Table 5 Macro-*F* scores of MaxEnt

主题	词特征	词特征+表情特征	词特征+主题词特征	全部特征
饮食	0.527	0.523	0.536	**0.544**
娱乐	0.316	0.339	0.356	**0.367**
购物	0.230	**0.257**	0.235	0.246
景观	0.382	0.367	**0.385**	0.383
交通	0.246	0.255	0.294	**0.327**
住宿	0.275	0.298	0.294	**0.336**
平均	0.329	0.340	0.350	**0.367**

说明：加粗数字为该主题下 Macro-*F* 的最大值，下同。

表 6 支持向量机 Macro-*F*值

Table 6 Macro-*F* scores of SVM

主题	词特征	词特征+表情特征	词特征+主题词特征	全部特征
饮食	0.482	0.485	0.493	**0.509**
娱乐	0.284	0.292	0.300	**0.327**
购物	0.192	**0.211**	0.176	0.192
景观	0.336	**0.341**	0.321	0.339
交通	0.225	0.222	0.246	**0.263**
住宿	0.240	**0.266**	0.237	0.265
平均	0.293	0.303	0.296	**0.316**

① https://github.com/fxsjy/jieba https://github.com/scikit-learn

大熵模型和支持向量机的 Macro-*F* 均值分别提高到0.367和0.316。按照主题来观察建模效果可以发现，同时引入表情特征和主题词特征时，最大熵模型在 4 个主题取得最优效果，支持向量机在 3 个主题取得最优效果。以上结果表明，引入表情特征和主题词特征可以提升建模效果，同时使用这两个特征效果达到最优。这也说明，表情特征和主题词可以有效地反映游客微博的主题情感倾向。

图 1 进一步分主题对比最大熵模型和支持向量机的建模效果。可以看出，两个模型对不同主题的分类效果总体变化规律较为一致，Macro-*F* 值从大到小依次为饮食、景观、娱乐、住宿、交通和购物。同时，在各个主题中，应用不同特征组合时，最大熵模型建模效果均优于支持向量机。

由于游客微博主题情感分析研究成果较少，未发现能够与本文进行直接定量对比的研究成果。本文选择与话题情感分析研究成果进行建模效果对比。文献[18]中列举了12个话题情感分析建模的研究成果，本研究中最大熵模型取得的Macro-*F*均值为0.367，优于文献[18]列举研究成果的Macro-*F*均值0.273，介于第 3 名NUSTM的Macro-*F*值0.382与第4名 CUCSAS的Macro-*F*值0.340之间，说明本文的总体建模效果良好。

4　结语

本文将机器学习方法引入游客微博主题情感分析的研究工作中。相对于既有研究中常见的基于情感词典的计算方法，机器学习方法不依赖主观经验，可以对建模效果进行定量评价，并支持分主题输出情感倾向信息，可以为旅游管理和决策提供更直接和可靠的参考依据。本研究以赴日游客微博数据为实例，应用两种机器学习模型开展游客微博主题情感的分析研究，并对比 3 种特征对建模效果的影响。实验结果显示，这两种模型均能够有效地应用于主题情感分析，其中最大熵模型的效果略优于支持向量机；同时，在词特征的基础上，引入微博表情和主题词特征能够进一步提高建模效果。后续研究中，拟进一步分析和挖掘主题情感的表征方法，提高建模效果。

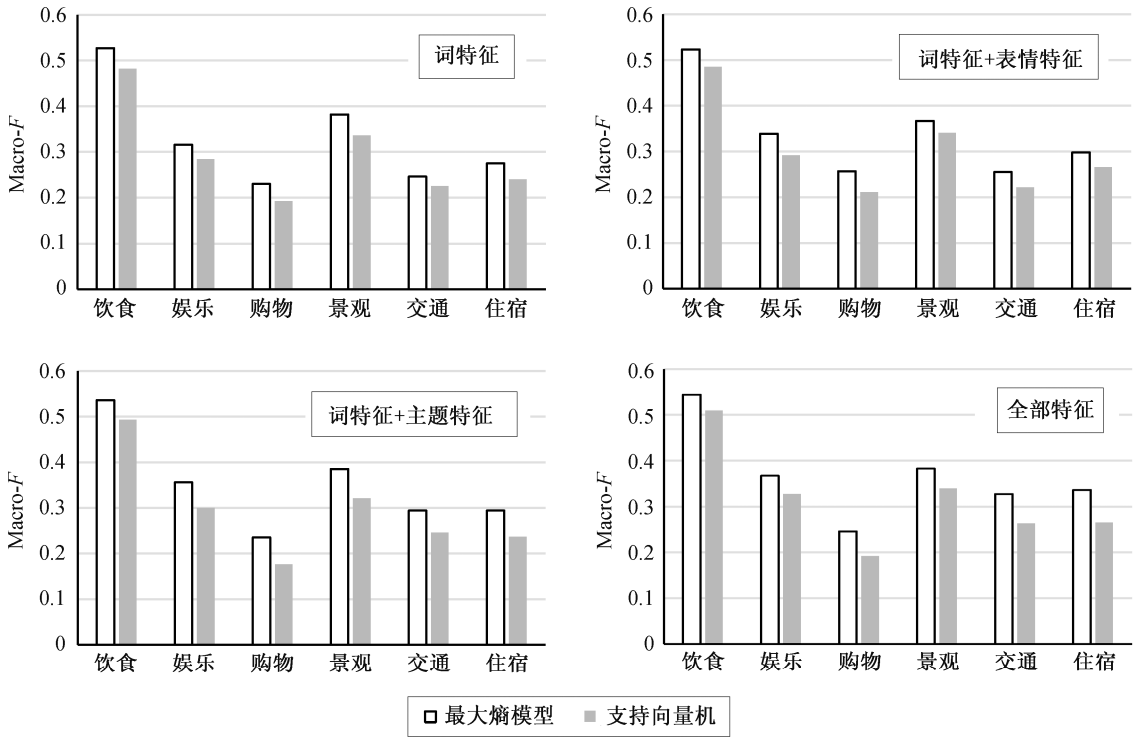

图 1　不同特征组合下各主题建模效果对比

Fig. 1　Comparison of Macro-*F* scores based on different feature sets in six themes

参考文献

[1] Bai X, Chen F, Zhan S B. A study on sentiment computing and classification of Sina Weibo with Word2vec // IEEE International Congress on Big Data. Anchorage, AK, 2014: 358–363

[2] Tse R T. Application of data mining in Sina Weibo — sentiment indicator to gauge tourist satisfaction in Macao. International Journal of Innovation, Management and Technology, 2016, 7(2): 80–85

[3] 周胜臣, 瞿文婷, 石英子, 等. 中文微博情感分析研究综述. 计算机应用与软件, 2013, 30(3): 161–164

[4] 周立柱, 贺宇凯, 王建勇. 情感分析研究综述. 计算机应用, 2008, 28(11): 2725–2728

[5] 涂海丽, 唐晓波. 基于在线评论的游客情感分析模型构建. 现代情报, 2016, 36(4): 70–77

[6] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Engineering Journal, 2014, 5(4): 1093–1113

[7] 中国旅游研究院. 中国出境旅游发展年度报告 2016. 北京: 旅游教育出版社, 2016: 3–4

[8] 张思豆, 李君轶. 基于微博大数据的游客情感与空气质量关系研究——以西安市为例. 陕西师范大学学报(自然科学版), 2016, 44(4): 102–107

[9] Girardin F, Calabrese F, Fiore F D, et al. Digital footprinting: uncovering tourists with user-generated content. IEEE Pervasive Computing, 2008, 7(4): 36–43

[10] Xiang Z Q, Zou Y X, Wang X. Sentiment analysis of Chinese micro-blog using vector space model // Signal and Information Processing Association Annual Summit and Conference (APSIPA). Siem Reap: IEEE, 2014: 1–5

[11] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques // ACL-02 Conference on Empirical Methods in Natural Language Processing. Philadelphia, 2002: 79–86

[12] 曹宇, 王名扬, 贺惠新, 等. 情感词典扩充的微博文本多元情感分类研究. 情报杂志, 2016, 35(10): 185–189

[13] 张珊, 于留宝, 胡长军. 基于表情图片与情感词的中文微博情感分析. 计算机科学, 2012, 39(增刊 3): 146–148

[14] 刘宝芹, 牛耘, 张景. 基于统计数据的微博表情符分析及其在情绪分析中的应用. 计算机工程与科学, 2016, 38(3): 577–584

[15] 石晶, 李万龙. 基于 LDA 模型的主题词抽取方法. 计算机工程, 2010, 36(19): 81–83

[16] Ye D S, Huang P J, Hong K D, et al. Chinese microblogs sentiment classification using maximum entropy // 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and 7th International Joint Conference on Natural Language Processing. Beijing, 2015: 171–179

[17] Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2002, 2(1): 45–66

[18] Liao X W, Li B Y, Xu L H. Overview of Topic-based Chinese Message Polarity Classification in SIGHAN // 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and 7th International Joint Conference on Natural Language Processing. Beijing, 2015: 56–60