计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):118−123 [doi: 10.15888/j.cnki.csa.006498]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于词向量技术与主题词特征的微博立场检测①

郑海洋, 高俊波, 邱　杰, 焦　凤

(上海海事大学 信息工程学院, 上海 201306)
通讯作者: 郑海洋, E-mail: zhenghy53@163.com

摘　要: 微博话题随着移动互联网的发展变得火热起来, 单个热门话题可能有数万条评论, 微博话题的立场检测是针对某话题判断发言人对该话题的态度是支持的、反对的或中立的. 本文一方面由 Word2Vec 训练语料库中每个词的词向量获取句子的语义信息, 另一方面使用 TextRank 构建主题集作为话题的立场特征, 同时结合情感词典获取句子的情感信息, 最后将特征选择后的词向量使用支持向量机对其训练和预测完成最终的立场检测模型. 实验表明基于主题词及情感词相结合的立场特征可以获得不错的立场检测效果.

关键词: 立场检测; 主题词特征; 词向量; 立场特征

引用格式: 郑海洋,高俊波,邱杰,焦凤.基于词向量技术与主题词特征的微博立场检测.计算机系统应用,2018,27(9):118–123. http://www.c-s-a.org.cn/1003-3254/6498.html

Stance Detection in Chinese Microblog Topic Based on Word Embedding Technology and Thematic Words Feature

ZHENG Hai-Yang, GAO Jun-Bo, QIU Jie, JIAO Feng

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: With the development of the mobile Internet, Microblog topic has become popular. A single hot topic may have tens of thousands of comments. The stance detection of Microblog topic aims to automatically determine whether the author of a text is in favor of the given target, against the given target, or neither. Firstly, Word2Vec trains out each word of the corpus of vector to extract semantics information from sentence. Then, TextRank keywords extracted method is used to construct the thematic words set as the stance's feature, meanwhile, the sentiment lexicon is used to extract the sentiment information of the sentence. Finally, the word vector of feature selection is trained and predicted by Support Vector Machine (SVM), so as to complete the model of stance detection. The experimental result shows that the stance feature based on the combination of thematic words and sentiment words can obtain good stance detection effect.

Key words: stance detection; thematic words feature; word embedding; stance feature

1　引言

据统计, 微博客户端日活跃用户数达 1.54 亿, 用户不仅可以在微博平台上上传图片记录自己的生活, 也可以通过转发、点赞、评论等与其他用户进行互动, 或针对某一话题公开发表自己的立场及观点. 立场检测 (stance detection) 是通过微博作者针对某一话题发表的评论, 检测微博作者对该话题的立场是支持、反对或中立, 这一检测结果不仅可以帮助政府了解民情、完善法律法规, 而且可以帮助商业公司对产品功能进行改进, 提高用户体验. 近年来, 这一课题吸引了众多学术界及工业界的关注和研究, 并成为自然语言处理 (Natual Language Processing, NLP) 中的一个新兴热门研究领域[1,2].

微博话题的立场检测与传统的文本情感分析非常

① 收稿时间: 2018-01-06; 修改时间: 2018-01-23; 采用时间: 2018-02-01; csa 在线出版时间: 2018-08-16

相似, 但又有着明显的区别. 单纯的使用微博的情感信息并不能把握微博作者的立场, 如微博“这些骑电车横冲直撞的人太可恶了, 上次就差点被撞到, 太危险了”, 此微博所抒发的是负面情绪, 但针对“深圳禁摩限电”这个话题表明的态度却是支持的. 通常微博作者在表达某个话题的立场倾向时, 往往会说出自己支持或反对的理由, 如针对上述话题, 支持者所表达的核心一般都是以交通安全为主题, 而反对者的理由通常是一些普通老百姓出行不方便所带来的一系列问题, 如果在该话题下能提取出到网民支持或反对的核心理由, 就能准确地判断发言人的立场倾向.

随着 Bengio 等人提出了神经网络语义模型之后, 文本词向量技术受到了广泛关注. 对于机器学习而言, 特征表示将直接影响模型性能的好坏, Word2Vec 模型可以根据文本中词的关联性有效地学习语义特征. 本文结合话题文本构建出主题词集, 在词向量下根据主题词和情感词筛选出有价值的特征信息, 最后使用 SVM (Support Vector Machine) 对话题语料进行训练得到最终的立场检测模型.

2 相关工作

微博立场的检测属于文本倾向性分析的研究领域, 文本倾向性分析又主要分为情感分析和立场分析. 常见的情感分析方法也可以适用于立场分析, 目前针对微博文本的情感分析, 近年来已取得了非常出色的研究成果. 2011 年, Lu 等人[3]提出了一种基于统一规则的自动化构建情感词典的方法, 该方法针对情感分析表现不错, 但不能提取到文本的隐含语义信息. 针对立场分析, Ebrahimi 等[4]将情感极性融入到对象和立场中, 并对三者进行对数线性联合建模. Chen 等[5]通过话题风格和微博文本进行特征提取, 建立了基于 CNN 的立场检测模型. Liu 等[6]提出了一种以情感加权算法和朴素贝叶斯算法相结合的组合分类模型, 该模型虽有较好的立场判别精度, 但并不能处理复杂的中文句式及上下文语境等情况. Dian 等[7]通过探究不同的特征提取方法, 之后使用支持向量机、随机森林和梯度提升决策树对上述特征进行立场检测, 最后结合所有的特征分类器进行后期融合, 在 2016 年自然语言处理与中文计算会议 (NLPCC2016) 中文微博立场检测评测任务中取得了第一名的成绩. 随着以 Word2Vec 为代表的词向量广泛应用, 本文基于词向量技术并构建主题词集作为立场特征完成微博话题的立场检测.

3 基于主题词特征的微博话题立场检测

3.1 模型框架

如图 1 所示, 本文主要使用 Word2Vec 和立场特征对微博进行立场检测. 首先对数据集进行数据清洗, 同时构建当前话题的主题词集. 然后使用词向量技术对所有词汇进行学习获得其词向量表示, 接着使用三种方法对每条微博做特征选择, 之后将筛选后的词向量求平均值作为每条微博的最终特征向量, 最后采用 SVM 算法对上述特征向量进行训练和预测得到最终立场检测模型.

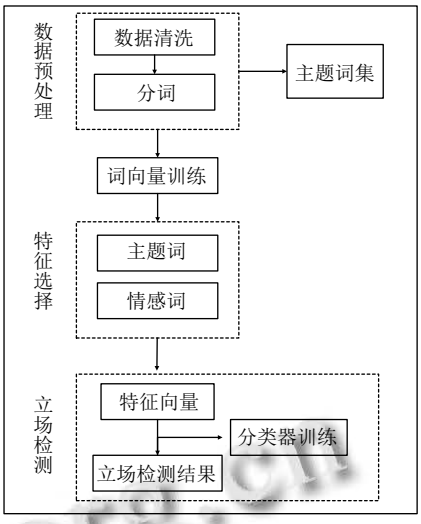

图 1 微博话题立场检测模型

3.2 文本预处理

文本预处理[8]包括数据清洗、分词、去停用词. 数据清洗主要过滤掉微博的文本噪音, 其中主要去除以“#”包围的字段、url 网址、表情符号、转发符号, 本文采用正则表达式去除上述噪音. 之后对文本进行分词, 分词工具采用的是中文开源分词工具 jieba 分词. 最后去除停用词, 去除停用词可以对文本进一步降噪, 停用词主要包含一些标点符号、代词、助词, 如“他”、“要”、“也”、“。”等.

3.3 主题词集

常用的主题词集构建方法有 TF-IDF 词频逆文档频率和 TextRank 算法[9]. TF-IDF 是基于词袋模型 (Bag-of-Words), 通常把文章表示成词汇的集合, 而不考虑句子中词汇的顺序关系, 不能有效地反应文章的

内部组织结构. TextRank 算法类似于 PageRank 算法, 它是基于网络模型 (graph model), 该模型将文章表示成网络结构, 网络中各个节点表示单个词汇, 节点之间的边表示文章中词汇之间的邻近关系, 该方法能联系到句子中词汇的位置关系. 由于微博句式简单, 结构相似, 为了获取话题的核心关键词, 本文采取 TextRank 算法构建主题词集. 如表 1 所示, 列举了 TextRank 提取的部分主题词.

表 1 TextRank 提取的部分主题词

部分主题词集
{‘放鞭炮’, ‘鞭炮’, ‘传统’,‘大家’, ‘爆竹’,‘燃放烟花’, ‘烟花爆竹’,‘烟花’, ‘空气’, ‘市民’, ‘孩子’,‘民俗’, ‘文化’, ‘习俗’,‘倡议’, ’回家’, ‘传统节日’, ‘空气质量’,‘鞭炮声’, ‘垃圾’, ‘方式’, ’饺子’, ‘环境’, ‘环卫工人’,‘烧纸’, ‘建议’, ‘燃放鞭炮’, ‘味儿’, ‘家人’,‘绿色’, ‘放炮’, ‘问题’,‘火灾’,‘污染环境’, ‘天气’, ‘父母’,‘汽车’,‘新衣’}

3.4 词向量模型

在文本分类中, 特征抽取通常根据某个特征评估函数计算每个特征的评分值, 并以此作为权重按评分值进行排序, 然后选取若干个评分最高的作为特征词. 这种类型的算法有词袋模型 (Bag-of-Words-BOW) 算法, N-Gram 算法, 但这类算法无法提取词汇的深度语义信息. Bengio 等人基于 n-gram 思想提出了神经网络语言模型[10], 采用三层神经网络学习词向量[11], 其核心思想是常用的神经网络算法, 该模型能有效地提取文本的深度特征, 近年来已经有很多学者使用该技术做情感分析[12], 并取得了非常出色的实验效果.

Word2Vec 根据语料库的词汇顺序关系, 利用 CBOW 模型或 Skip-Gram 模型将词汇转换为 K 维空间的向量表示 (distributed representation)[13], 模型包含输入层、投影层和输出层. 其中 CBOW 模型是根据当前词 w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} 的前提下预测 w_t, 而 Skip-Gram 模型恰恰相反. 本文采用的是 CBOW 词向量模型, 其基于神经网络语言模型的目标函数通常取如下对数似然函数:

$$L=\sum_{w\in C}\log p(w|Context(w)) \tag{1}$$

如图 2 所示, CBOW 词向量模型由三层神经网络构成, 其中输入层包含 $Context(w)$ 中 $2c$ 个词向量, $v(Context(w)_1),\cdots,v(Context(w)_{2c})\in R^m$, 由词 w 前后各 c 个词构成, m 表示词向量的长度, 投影层将输入层 $2c$ 个词向量做求和运算, 公式如下:

$$x_w=\sum_{i=1}^{2c}v(Context(w)_i) \tag{2}$$

输出层根据每个词在语料库中出现次数构造一颗 Huffman 树, 叶子节点分别为词典 D 中的每个词, 一共有 N 个, 非叶子结点有 N–1 个. Word2vec 将 Huffman 编码为 0 的节点定义为正类, 编码为 1 的节点定义为负类, 这样对于词典 D 中任意一个词 w, Huffman 树一定有一条从根结点到词 w 对应结点的唯一路径 p^w. 路径 p^w 上存在 l^w-1 个分支, 每个分支上对应一个二元分类的概率, 将这些概率连乘的结果就是最后所需的条件概率 $p(w|Context(w))$, 计算公式如下:

$$p(w|Context(w))=\prod_{j=2}^{l^w}p(d_j^w|x_w,\theta_{j-1}^w) \tag{3}$$

$$p(d_j^w|x_w,\theta_{j-1}^w)=\begin{cases}\sigma(x_w^{\mathrm{T}}\theta_{j-1}^w), & d_j^w=0\\ 1-\sigma(x_w^{\mathrm{T}}\theta_{j-1}^w), & d_j^w=1\end{cases} \tag{4}$$

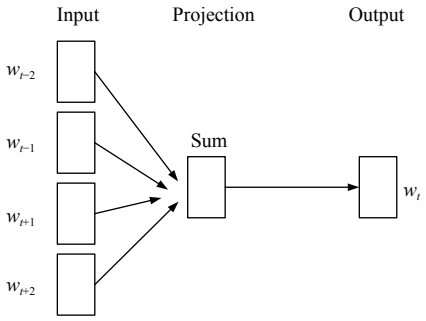

图 2 CBOW 词向量模型

本文利用 Gensim[14] 工具对语料中的词汇做 Word2Vec 训练, 词向量维数设置 500 维, 训练完后得到最终词向量 CBOW 模型. 词向量的相似性通常选取余弦距离来衡量, 两个词向量的余弦距离表示其在空间上的向量夹角, 余弦值越接近于 1 表明其夹角越接近 0 度, 也说明两个词向量越相似, 余弦距离计算公式如下:

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{5}$$

如表 2 所示, 在话题“春节放鞭炮”下训练的词向量模型得与词“放鞭炮”余弦距离最近的 5 个词语. 由余弦距离可以看出较近的词语通常代表的立场也相似, 其表明了微博作者持该立场的理由, 如“放鞭炮”与“环

境”、“污染”、“雾霾”相近, 说明该话题下针对“放鞭炮”谈论的核心是与环境相关的问题.

表 2 与“放鞭炮”词向量相近的词语

相近词语	词向量的余弦距离
春节	0.422 227 412 462 2345
环境	0.376 391 470 432 2815
雾霾	0.332 929 044 961 9293
污染	0.332 424 759 864 807 13
环保	0.318 536 818 027 496 34

3.5 特征选择

情感词和主题词通常决定微博作者的情感倾向和立场倾向, 更能表明微博作者的观点. 本文分别采用情感词和主题词的特征选择方法提取其立场特征, 情感词典综合采用中国知网情感词典、清华大学李军中文褒贬义词典和台湾大学 NTUSD 简体中文情感词典, 主题词集由 TextRank 算法在该话题下提取的 250 个名词构成. 如表 3 所示, 对预处理后微博分别采用主题词和情感词做特征选择之后的结果.

表 3 对微博文本分别采用两种特征选择的结果

预处理后的微博	情感词做特征选择	主题词做特征选择
{‘鞭炮声’, ‘忍耐’, ‘反对者’, ‘没有’, ‘差’, ‘惋惜’, ‘不能’, ‘气氛’, ‘放鞭炮’, ‘鞭炮’}	{‘忍耐’, ‘反对者’, ‘没有’, ‘差’, ‘惋惜’, ‘不能’}	{‘鞭炮’, ‘鞭炮声’, ‘放鞭炮’, ‘气氛’}
{‘空气’, ‘汽车’, ‘好’, ‘尾气’, ‘放鞭炮’, ‘提倡’, ‘喜欢’}	{‘好’, ‘提倡’, ‘喜欢’}	{‘汽车’, ‘放鞭炮’, ‘尾气’, ‘空气’}

3.6 模型的训练与预测

根据以往实验表明对于文本分类支持向量机与其他分类算法相比在处理非线性及高维分类中有着较好的分类效果[15]. 本文首先根据情感词典和主题词集从训练文本中获取其特征词语, 并利用词向量模型将其转换为 500 维的空间向量, 然后对其求平均值并做归一化处理作为最终的文本向量, 最后由 SVM 分类器根据训练集中的文本向量和相对应的正负标签训练出立场检测模型, 并利用生成后的模型对测试数据进行正负性的立场预测.

4 实验与分析

4.1 实验数据

本文采用的数据集来自 NLPCC2016 中文微博立场检测的评测任务, 选取的话题是任务中第二个话题“春节放鞭炮”, 该话题一共有 500 条微博, 其中持支持和反对立场的各有 250 条微博数据, 训练集与测试集比例为 8:2, 数据格式为微博 ID, 话题, 微博文本, 立场标签, 例如:

<ID>1

<TARGET>春节放鞭炮

<TEXT>又是一年新春到, 除夕之夜少放鞭炮, 为纯净空气做一份贡献.

<STANCE>AGAINST

4.2 评价标准

常用的评价标准通常选取准确率、召回率和 $F1$ 值, 本文针对立场检测为了综合考虑分类效果增加了正负类的 $F1$ 评价均值 F_{avg} 作为综合评价指标, F_{avg} 计算公式如公式 (6) 所示, 其中 F_{favor} 和 $F_{against}$ 分别表示支持立场和反对立场的 $F1$ 值, 本文实验中选取的话题“春节放鞭炮”在 NLPCC2016 比赛中最好的成绩 F_{avg} 是 77.61%.

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \tag{6}$$

4.3 实验结果与分析

为了获取更多的立场特征, 本文在对微博文本基于主题词和情感词特征选择后, 还将两种特征选择方法结合起来实验, 最后将筛选后的特征词转换为词向量求均值由 SVM 支持向量机对其训练及预测得到该话题下的最终立场检测模型, 实验结果如表 4 所示, 其中正类表示支持倾向, 负类表示反对倾向.

表 4 三种特征选择方法对对比实验效果

特征提取	正/负类	准确率 (%)	召回率 (%)	$F1$ 值 (%)	$F1$ 均值 (%)
主题词集	正类	77.77	84.00	80.76	79.96
	负类	82.60	76.00	79.16	
情感词典	正类	76.59	72.00	74.22	74.97
	负类	73.58	78.00	75.72	
{主题词集, 情感词典}	正类	81.64	83.56	82.60	82.59
	负类	81.45	83.73	82.59	

由表 4 所示, 在以 $F1$ 均值这项评价指标下, 以情感词典作为筛选特征在三种分类模型中表现最差, 说明传统的情感词典方法并不完全适合立场检测. 因为用户通常在表达立场时, 有时会包含自己的主观情绪, 但这种情绪具有两面性, 针对的可能是话题的正面, 也可能是话题的对立面, 所以情绪并不能准确地反映作者的立场倾向. 而基于主题词集的特征选择方法效果更好, 因为在话题中这些主题词代表的是该话题的核

心关键词, 也是微博作者支持或反对理由的主要理由, 更能反映发言人的论点和立场信息.

在使用主题词和情感词两种特征结合实验, 分类效果达到最好. 这是因为一些短微博往往无法提取到主题词, 但是它们多数含有一些情感词, 如果结合情感词做特征选择可以弥补主题词特征选择方法的弊端. 其次, 本文还针对主题词的数量做了定性实验, 在区间[100, 600]分别设置 7 种不同的主题词数量完成优化实验, 实验结果如图 3 所示.

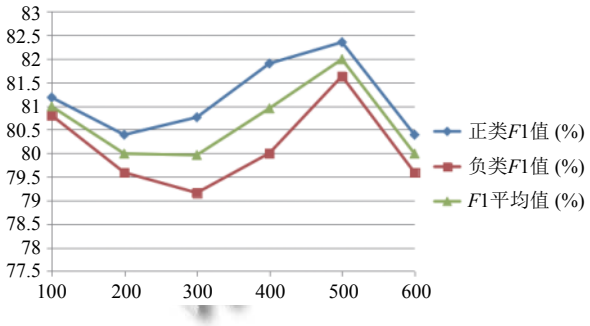

图 3 不同主题词个数的实验效果

从图 3 中可以看出选取 500 个主题词可以达到最好的立场检测效果, 因为当主题词较少时, 相对的立场特征就少. 相反, 当主题词过多时, 便会产生一些噪声干扰检测. 本文在选取 500 个主题词同时结合情感词典做辅助特征选择, 为了获得最好的分类效果, 通过调整 SVM 算法的惩罚系数 C, 实验结果显示在 C=5 时正负 $F1$ 均值可以达到 83%, 相比该话题最好的成绩提高了 5 个百分点, 证明本文方法的有效性.

从实验可以看出, 基于词向量技术和主题词特征在中文微博立场检测中可以获得不错的分类效果, 主要原因是 Word2Vec 训练出的词向量包含了词汇之间的语义信息, 其相对于词频特征更能表明词语的实际意义. 同时基于主题词特征的特征选择方法可以获取更多有价值的立场特征, 针对一些较短的微博, 结合情感词进一步提升了模型的性能.

5 结束语

本文通过使用 NLPCC2016 中文微博立场检测的数据集进行实验, 首先将文本进行预处理并使用词向量技术将词汇转换为高维空间的向量表示, 然后使用 TextRank 提取话题的关键词作为话题的主题词集, 提出了基于主题词的特征选择方法, 同时融入情感词典做辅助特征选择, 并使用支持向量机对话题微博进行训练及预测. 实验结果表明, 本文的方法在中文微博话题下具有较好的立场检测效果.

鉴于微博平台下话题众多, 如果进一步考虑话题的类别, 本文的研究方法还需做进一步扩展. 其次, 本文最后使用的文本向量是由特征选择后的词向量求均值所得, 这种方法虽然可行但丢失了词向量的顺序信息. 但是, 词向量隐含地包含了词汇间的顺序关系, 本文最后输入到 SVM 的特征向量是由特征选择后的词向量求均值所得, 它综合了当前微博的所有立场特征信息, 是当前微博立场的一般特征表示, 对微博立场检测的结果影响不大, 但具有研究价值. 在今后的研究中, 针对话题类别和如何获取句子的结构化信息是本文的研究重点.

参考文献

1 Anand P, Walker M, Abbott R, *et al*. Cats rule and dogs drool!: Classifying stance in online debate. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Stroudsburg, PA, USA. 2011. 1–9.

2 Walker MA, Anand P, Abbott R, *et al*. Stance classification using dialogic properties of persuasion. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA. 2012. 592–596.

3 Lu Y, Castellanos M, Dayal U, *et al*. Automatic construction of a context-aware sentiment lexicon: An optimization approach. Proceedings of the 20th International Conference on World Wide Web. New York, NY, USA. 2011. 347–356.

4 Ebrahimi J, Dou DJ, Lowd D. A joint sentiment-target-stance model for stance classification in tweets. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan. 2016. 2656–2665.

5 Chen WF, Ku LW. UTCNN: A deep learning model of stance classificationon on social media text. Proceedings of 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan. 2016.

6 刘勘, 田宁梦, 王宏宇, 等. 中文微博的立场判别研究. 知识管理论坛, 2017, 2(3): 175–185.

7 奠雨洁, 金琴, 吴慧敏. 基于多文本特征融合的中文微博的立场检测. 计算机工程与应用, 2017, 53(21): 77–84. [doi: 10.3778/j.issn.1002-8331.1702-0292]

8 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, 181. [doi: 10.3969/j.issn.1000-386x.2013.03.043]

9 顾益军, 夏天. 融合 LDA 与 TextRank 的关键词抽取研究. 现代图书情报技术, 2014, (7–8): 41–47.

10 Bengio Y, Réjean D, Vincent P, *et al*. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3(6): 1137–1155.

11 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Computer Science, 2013: 1–12.

12 Xue B, Fu C, Zhan SB. A study on sentiment computing and classification of Sina Weibo with Word2Vec. IEEE International Congress on Big Data. Anchorage, AK, USA. 2014. 358–363. [doi: 10.1109/BigData.Congress.2014.59]

13 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 2013. 3111–3119.

14 Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Malta. 2010. [doi: 10.13140/2.1.2393.1847]

15 郭丽娟, 孙世宇, 段修生. 支持向量机及核函数研究. 科学技术与工程, 2008, 8(2): 487–490. [doi: 10.3969/j.issn.1671-1815.2008.02.041]