HITSCIR_Run：COAE2015 微博观点句识别任务分析系统*

李泽魁[1]，赵妍妍[2]，秦兵[1]，刘挺[1]

[1]哈尔滨工业大学计算机科学与技术学院信息检索研究中心，哈尔滨，150001

[2]哈尔滨工业大学机电学院媒体系，哈尔滨，150001

E-mail: {zkli, yyzhao, qinb, tliu}@ir.hit.edu.cn

摘　要：文本倾向性分析作为自然语言处理领域研究的热点问题之一，目的是通过文本来分析人们的情感倾向。本文主要研究第七届中文倾向性分析评测（COAE 2015）中的微博观点句识别任务（任务 2），通过词典资源收集、词向量训练、特征抽取模板设计、训练语料构建、分类模型调优与集成等步骤，完成了一套有监督学习的基于特征抽取的情感倾向性分析系统。本系统在评测任务中取得了较好的成绩。

关键词：情感倾向性分析；特征抽取；情感专属词向量；模型集成

HITSCIR_Run: Sentiment Analysis System of Micro-blog Opinion Recognition Task in COAE2015

Zekui Li[1], Yanyan Zhao[2], Bing Qin[1], Ting Liu[1]

[1]Research Center for Social Computing and Information Retrieval of Computer Science and Technology School, Harbin Institute of Technology, Harbin 150001

[2] Department of Media Technology and Art, Harbin Institute of Technology, Harbin 150001

E-mail: {zkli, yyzhao, qinb, tliu}@ir.hit.edu.cn

Abstract: Text sentiment orientation analysis is a hot research feild in natural language processing. It aims to annalyze people's subjective opinions through plain text. This paper mainly focuses on the seventh Chinese Opinion Analysis Evaluation (COAE 2015) in micro-blog opinion recognition task (Task 2). We finally build a sentiment analysis system based on feature extraction algorithm, which is a kind of supervised learning method. More in detail, we did lots of works in training corpus and lexicon resources building, word embedding training, template of feature extraction design, model optimization and ensemble and so on. Our system achieves good results in the evaluation task.

Keywords: sentiment orientation analysis; feature extraction; sentiment specified word embedding; model ensemble

1　引言

情感分析又称意见挖掘，旨在研究人们针对实体、人物、事件、主题及其属性的主观意见和情感[1,2,3]。按照处理文本粒度的不同可以将情感分析分为篇章级[4]、句子级[5,6]、短语级[7]和词语级[8]。前五届中文倾向性分析评测（COAE）持续在词语级、句子级、篇章级进行了中文倾向性分析评测，任务涉及主客观分析、情感极性分析、评价对象抽取以及搭配抽取[9]等方面，为中文倾向性分析提供了一个很好的平台。为了探索中文倾向性分析的新技术与方法，推动中文倾向性分析理论和技术的研究与应用，本次评测共包含 4 个任务：

*基金项目：本课题受到国家自然科学基金重点项目(61133012)、国家自然科学青年基金项目(61300113)及国家自然科学基金面上项目(61273321) 资助。

- 任务 1：基于上下文的观点信息识别
- 任务 2：微博观点句识别
- 任务 3：微博观点句评价要素识别
- 任务 4：垃圾观点句识别

哈尔滨工业大学社会计算与信息检索研究中心开发了 HITSCIR_Run 分析系统参加了 COAE2015 的任务 2，即微博观点句识别任务。即给定微博集合，判别每个文本的情感倾向性，分为褒义倾向、贬义倾向和褒义贬义混合倾向，是一种句子级的情感分析任务。

本文内容安排如下：第二部分介绍本任务中所用到的资源收集工作，主要分为所用情感词典资源和情感专属词向量资源；第三部分对任务中用到的特征抽取模板进行详细说明；第四、第五部分阐述了实验部分，即训练语料的构建、模型的调优与集成和最终的评测结果。最后一部分对本文进行总结。

2 实验资源构建

2.1 词典资源收集

本文在进行特征抽取模板设计之前，首先收集了一系列可能会用到的词典资源。词典资源，分为情感词典（例如褒贬词典）和非情感词典（例如副词词典、否定词词典等）两种。词典资源作为情感分类任务的一项不可或缺的前提，在各类情感分析任务中起着至关重要的作用。

2.1.1 情感词典资源收集

实验中我们用到的国内诸多优秀的褒贬情感词典资源，表 1 简要介绍了这些词典的相关信息。

表 1 国内优秀情感分类词典资源举例

Tab.1 Valuable lexicon resources of Chinese sentiment analysis

ID	词典名称	创作版权	情感分值	词典规模
1	哈工大褒贬词典	哈尔滨工业大学 社会计算与信息检索研究中心	无	21690
2	大连理工褒贬词典	大连理工大学信息检索研究室	有	22012
3	台湾大学 NTUSD	台湾大学	无	11086
4	清华大学褒贬词典	清华大学李军等人	无	10035
5	清华大学情感词典	清华大学原博等人	有	23419
6	哈工大大规模情感词词典	哈尔滨工业大学 社会计算与信息检索研究中心	无	150000

其中，词典 1 由哈尔滨工业大学社会计算与信息检索研究中心（HIT-SCIR）文本挖掘组编写；词典 2 基于大连理工大学信息检索研究室（DUTIR）的情感词汇本体库[10]，通过筛选出具有明显极性的词语与对应的极性两个维度的数据，同时做了细微处理来用于情感分类；词典 3 是台湾大学的意见词词典，又名 NTUSD，具体编写人员暂时无法考证；词典 4 为清华大学李军等人构建的情感词褒贬词典[11]；词典 5 为清华大学中文系原博构建的情感词典，比较遗憾，该词典尚未有具体构建算法的论文发表；词典 6 为哈尔滨工业大学社会计算与信息检索研究中心（HIT-SCIR）赵妍妍等人使用大规模弱标注情感分类算法自动标注微博语料，从中自动构建出的总数 15W 条的大规模情感词词典[12]。值得一提的是，带有具体情感强度的词典为词典 2 和词典 5。

2.1.2 非情感词典资源收集

情感分类任务除了依赖一定的情感词典资源，还离不开一些影响句子情感极性的非情感词典资源。在本小节将非情感词典资源分为修饰词词典和连词词典两大部分。具体信息见表 2。

表 2 非情感词典资源举例

Tab.2 Chinese modified word dictionary and conjunction dictionary

词典资源类型	词典名称	词典规模	词汇样例
修饰词词典	否定词	107	不、别
	强程度副词	193	真的、最
	弱程度副词	45	稍微、有点
	疑问副词	15	哪里、有啥
连词词典	总结或转折连词	40	总之、但是

其中，修饰词词典包括了否定词、强弱程度副词和疑问副词三部分。在实际语料中我们发现，当情感词被这些修饰词围绕时，有很大概率伴随着整句的情感极性变化，例如极性反转、加强或减弱等。另一方面，连词词典作为判定句子结构的依据，例如总结性和转折性的连词会对长句的语义及情感倾向进行总结，抑或对前半句的情感极性进行翻转。本文会在第三章介绍非情感词典的使用场景。

2.2 情感专属词向量资源构建

深度学习在近几年来迅猛发展，日新月异。在情感分类领域，部分学者通过对文本的基本组成单元——词汇进行建模。Labutov 等人[13]在 2013 年提出的“词向量重构（Re-Embedding）”方法，是一种基于现有根据上下文信息计算出的词向量（Word Embedding），通过有监督地加入情感信息，使得重构出的词向量有更好的情感分类效果。国内青年学者 Tang 等人[14]另辟蹊径，他首先通过上文提及的根据表情符对微博文本自动弱标注的方法对海量微博标注情感，然后直接在训练情感专属词向量（SSWE，Sentiment Specified Word Embedding）时直接有监督地引入情感信息。

Tang 等人将 Collobert 等人[15]提出的 C&W 模型的输出层的维度更改为情感类别数，同时在原有的输出层上增加 softmax 层来预测 SSWE 模型的输出条件概率。不同于原先的无监督的腐蚀某一个词的方法，Tang 将损失函数定义为最小化预测的情感类别与实际的情感类别之间的差值，如公式 1 所示：

$$\text{Loss Function} = -\sum_{k=\{0,1\}} f_k^g(t) \cdot \log(f_k^h(t)) \tag{1}$$

其中，$f_k^g(t)$代表 k 情感类别的标准输出（Gold Result），$f_k^h(t)$代表 k 情感类别的预测输出。

本文使用了 Tang 等人封装的 SSWE-Hard 工具包，对情感专属词向量进行了训练。训练过程的相关信息如表 3 所示。

表 3 情感专属词向量训练信息表

Tab.3 Training information of SSWE

属性名称	属性值
训练输入	约 170 万句根据表情符自动弱标注的褒贬中三元情感的中文微博
训练输出	274789 条情感专属词向量，每个词语对应一个 50 维的浮点数向量

3 特征抽取模板设计

在进行特征抽取工作之前，本文首先通过哈尔滨工业大学社会计算与信息检索研究中心开发的语言云技术平台[16]（LTP, Language Technology Platform）对评测语料进行了分词与词性标注等预处理。同时对微博中的人名和 URL 等与文本情感极性无关的短语进行了必要的替换。

本文使用径向基核（RBF, Radial Basis Function）的支持向量机（SVM, Support Vector Machine）作为分类器算法，那么情感分类准确率的高低很大程度上取决于特征抽取的模板的设计的好坏。本文借鉴了 2015 年国际语义评测（SemEval, Semantic Evaluation）面向 Twitter 的情感分类任务的最优分类系统[17]，同时结合国内对中文微博情感倾向性分析的部分研究工作[18]，设计了如下特征抽取模板：

- N 元文法特征（N-gram 特征）

抽取时，本文采用了一元文法特征（Unigram 特征）和二元文法特征（Bigram 特征）。考虑到更高元的文法特征，例如三元、四元甚至五元文法特征的语言稀疏性，特征抽取时未将其列入模板中。

- 词典特征

这里使用了表 1 中列举的 6 个国内优秀的情感分类褒贬词典。通过对文本中的词语进行词典匹配，最终统计不同词典在一句文本的一些数值特征，例如褒义词数目、贬义词数目、褒贬义词数目差、文本中褒义词分值最大值和最后一个褒义词分值等特征。

- 多词典规则评分特征

本特征使用了表 2 中列举的非情感词典资源，如果相关修饰词在情感词文本周围一定窗口出现，那么对情感词的情感分值做如表 4 的处理。表 4 中的每一行对应一种语言现象，例如含有否定词、含有疑问词等。

表 4 多词典规则评分策略

Tab.4 Strategies of score calculation based on multi-dictionary and rules

微博文本示例	文本现象	评分策略
这道菜好吃。	情感词	+情感词分值
这道菜真好吃。	情感词+强程度副词	+情感词分值*2
这道菜还算好吃。	情感词+弱程度副词	+情感词分值*0.5
这道菜不怎么好吃。	情感词+短距离否定	-情感词分值
我不认为这道菜好吃。	情感词+长距离否定	-情感词分值
这道菜哪儿好吃了？！	情感词+短距离疑问	忽略情感词

通过这维特征的引入，可对情感倾向性判定起到一定的辅助作用。

- 情感专属词向量特征（SSWE）

本文使用了表 3 中的情感专属词向量的训练结果，通过统计句中每个词语的词向量的每个维度，追加每一维的池化（Pooling）信息，这里池化信息指简单的最大值、最小值和平均值信息。

- 词性特征

词性特征本文特指和句子中情感倾向相关的一些词性的出现频率，例如形容词、副词、标点等。

- 表情符特征

表情符，是微博文本特有的一种语言现象，同时是情感倾向的重要判别特征。虽然经过抽样统计，评测语料中含有表情符的微博只占很小一部分，本文仍然坚持使用了此特征。

- 否定词特征

这里提及的否定词特征是指文本中的“否定区间”内的词汇。所谓否定区间，本文将其定义为否定词到其后第一个标点之间的词区间。例如“这道菜不怎么好吃。”这段文本，它的否定区间是“怎么 好吃”，为了区别非否定区间，本文将其标为“怎么_NEG 好吃_NEG”，即“不”和句尾的标点之间的词语。

- 感叹词特征

统计文本中的感叹词数目、最后一个词是否是感叹词或者感叹型标点。

- 疑问词特征

统计文本中的疑问词数目、最后一个词是否是疑问词或者疑问型标点。

- 话题 Hashtag 特征

话题 Hashtag，也是微博文本特有的语言现象之一。用户一般将微博的话题信息写进两个“#”之间的区间，所以这里统计了是否有 Hashtag 、Hashtag 内文本的分词结果等信息。

4 实验设计

4.1 训练语料的构建

本次评测任务 2 给定的评测语料共计约 13 万句，为了了解语料主题构成，本文首先抽样部分数据进行了简单统计。发现语料中微博所占比例不是特别高，反而一些商品评论（例如相机、汽车、手机等领域）语料占了相当一部分比例。针对这个现象，本文收集了一套多主题混合的人工标注训练语料来学习分类模型。具体训练语料信息如表 5：

表 5 训练语料组成说明

Tab.5 Distribution of training corpus

ID	语料领域	极性分布	语料规模	标注单位
1	新浪微博	褒贬中	7019	哈工大社会计算与信息检索研究中心
2	“蒙牛”产品评论	褒贬	2278	COAE 2014 情感分类训练语料
3	新闻中性文本	中性	1000	哈工大社会计算与信息检索研究中心
4	相机领域产品评论	褒贬中	19752	哈工大社会计算与信息检索研究中心

从表 5 中我们不难发现，在训练集中本文刻意地增加了产品评论语料的比例，目的是尽量与评测语料主题上接近。

4.2 分类模型的调优与集成

训练语料经过人工标注构建完成后，接下来根据第三章设计的特征抽取模板进行模型训练，进而预测评测给定语料就是顺理成章的事情了。经过模型调优的步骤后，本文区别于传统的单分类模型预测，提出了多模型集成的分类算法，使得评测任务中所用到的分类模型更加鲁棒，取众家之长。

上面提及的多模型集成的分类算法，本任务在表 5 张提到的训练语料基础上，选择不同的特征组合训练了三个不同的分类模型（如表 6 的第 1-3 行）。同时，对不同的分类模型进行了排列组合的集成处理（如表 6 的第 4-6 行）。他们的特征抽取组合如表 6 所示。

表 6 多模型集成分类算法模型简介

Tab.6 Introduction of multi-model ensemble classification algorithm

ID	模型代号	特征抽取模板
1	Basic Features(BF)	除去 SSWE 特征外的其他基础特征
2	BF +SSWE	Basic Features +SSWE 特征
3	BF +SSWE_normalized	Basic Features +SSWE 归一化的特征
4	Ensemble(1&2)	模型 1 和模型 2 的集成模型
5	Ensemble(1&3)	模型 1 和模型 3 的集成模型
6	Ensemble(2&3)	模型 2 和模型 3 的集成模型

为了评估每个分类模型的性能，本文在训练语料上对模型的分类效果进行了交叉验证。在这里本文计算了褒义、贬义、中性三种分类结果的精确率，作为评价模型优劣的指标。其在训练语料上的分类结果如表 7 和图 1 所示。

表 7 模型分类效果对比

Tab.7 Accuracy comparison between classification models

ID	模型代号	褒义精确率（%）	贬义精确率（%）	中性精确率（%）
1	Basic Features(BF)	57.20	63.65	**65.14**
2	BF +SSWE	61.47	67.04	55.35
3	BF +SSWE_normalized	60.38	65.86	57.63
4	Ensemble(1&2)	**62.56**	**67.93**	62.14
5	Ensemble(1&3)	61.49	66.72	63.35
6	Ensemble(2&3)	61.34	67.17	55.89

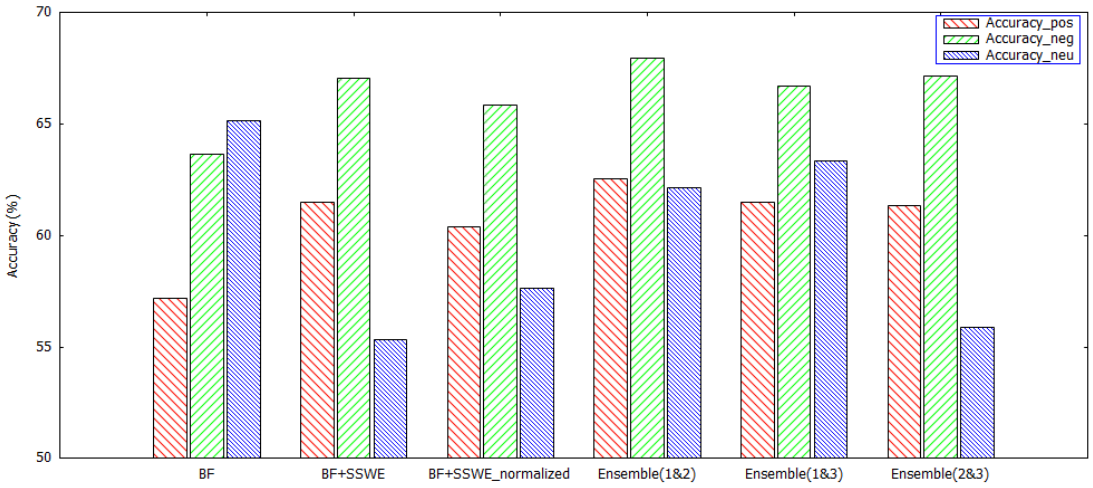

图 1 模型分类效果对比柱状图

Fig.1 Accuracy histogram of different classification models

表 7 中，某一个情感倾向的分类精确率（Accuracy）在本任务中的定义为某一类情感倾向的文本分类正确的数目与实际的相关文本数的比值。计算公式如公式 2：

$$\text{Accuracy}_{emotion} = \frac{\text{emotion 类别分类正确的文本数目}}{\text{emotion 类别文本总数}} \quad (2)$$

从三个普通模型的分类精确率结果上看，没有加入情感专属词向量特征的分类模型（模型 1），在中性语料上分类精确率高于加入情感专属词向量特征的模型（模型 2 和模型 3）。与之相反的是，后者在褒义贬义分类精确率上要略高于前者。褒义贬义分类精确率最高的模型出现在模型 1 与模型 2 的集成模型；中性分类精确率最高的是模型 1。

深入分析实验结果后我们发现，随着情感专属词向量特征的引入，由于引入更加丰富的词汇情感上下文信息，许多褒义贬义的分类难点（例如隐喻、讽刺等语法现象）更容易分类正确，例如下面的例子：

(1)“自从安装了八核 CPU，速度飕飕的。”

(2)“又买了一个碰地就碎的手机。。。”

上述两句中分别体现了隐喻和讽刺的语法现象，传统的基础特征抽取模板训练出的模型是很难将上述现象分类正确的。但是实验中我们发现，本模型可以保证对这类语法现象一定的分类精确率。

针对加入情感专属词向量特征后，模型对中性语料分类精确率降低的问题，通过观察分类错误对应的语料，我们发现问题主要集中在极短的中性文本上。例如下面的例子：

(1) 感觉咋样？

(2) 凯越销售额为 1500W 辆。

由于这部分例子实际上是不具有情感倾向的，但是文本虽然短，其亦可以映射一套特征向量，这样就增大了分类器的分类难度。这部分拉低了中性语料的分类精确率。

5 评测结果及分析

微博观点句识别任务的评价指标包括褒义、贬义和褒义贬义混合三种情感倾向的准确率（Precision）、召回率（Recall Rate）和 F 值（F-Score），以及三种评价指标的宏平均、微平均。COAE 评测举办方提供评测语料的标准情感倾向。表 8 为评测结果对比表格。

表 8 微博观点句识别任务评测结果

Tab.8 Result of Micro-blog Opinion Recognition Task in COAE2015

System	pos_F1	neg_F1	mix_F1	mirco_F1	marco_F1
HITSCIR_Run2	0.7562	0.7114	0.2236	0.6411	0.6106
HITSCIR_Run1	0.6596	0.6840	0.1993	0.5573	0.5821
Best	0.8617	0.7868	0.4233	0.8113	0.6945
Medium	0.7233	0.5800	0.1617	0.6057	0.4947

按照 4.2 节的模型集成调优的结果，本文最终向评测组委会提交了两套不同的模型分类预测结果。表 8 中第一行的 HITSCIR_Run1 使用的是表 6 中的模型 2（BF +SSWE），即分类模型特征抽取模板为在基础分类特征基础上追加情感专属词向量特征；第二行的 HITSCIR_Run2 使用的是模型 4（Ensemble1&2），即模型 1 和模型 2 的集成模型。由于模型 1 的中性语料分类精确率较高，而模型 2 的褒义贬义倾向的分类精确率较高，所以对模型 1 和模型 2 的分类投票权重做了偏执处理。最终，HITSCIR_Run2 分类预测结果由于前者，最终系统排名第 5 名，机构组织排名第 3 名。

经过对比最优系统（Taojin_0）的评测结果、仔细分析所提交的语料，本文发现了本系统的不足，问题集中在以下两点：

- 褒义贬义混合判别准确率不高

本系统的分类模型中加入了情感专属词向量，一方面提高了文本的情感分类的准确率，弥补了许多传统特征抽取方法无法分类准确的语言现象的空白；另一方面，它对中性语料，或者说是情感倾向极不明显的语料，更倾向于预测为带有情感的。后者同时降低了中性语料与非中性语料两部分预测结果的 F 值。

- 评测提交规则未理解透彻

通过计算本任务最优系统的评测结果，计算出了该系统最终提交预测结果数为 96897 句，占评测语料总数的 72.7%。遗憾的是，本系统对评测提交规则没有理解透彻，提交了所有 133201 句评测语料，导致某些不相关的中性文本也大幅拉低了评测结果。即使评测结果有这样的大幅缩水，本系统的分类效果仍然名列前茅，算法准确率和鲁棒性等方面还是可圈可点的。

6 总结

哈尔滨工业大学社会计算与信息检索研究中心所设计的 HITSCIR_Run 情感分类系统参与了 COAE2015 的微博观点句识别任务。本文提出了多模型集成的情感倾向性分析系统最终取得了不错的成绩，各个指标均接近该任务最优系统。具体的，在特征抽取模板设计部分，本文在情感分类基础特征模板基础上提出了多词典混合、复杂词典规则评分和情感专属词向量等创新特征；在模型调优与集成部分，本文通过交叉验证得出了最优的集成分类模型，最终的评测结果也证实了本文分类方法的有效性。

参 考 文 献

[1] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834-1848.

[2] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[3] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167.

[4] Tang D, Qin B, Liu T. Learning Semantic Representations of Users and Products for Document Level Sentiment Classification[C]// Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics, 2015: 1014-1023.

[5] Zhu X, Kiritchenko S, Mohammad S M. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets[J]. SemEval 2014, 2014, 443.

[6] Rosenthal S, Nakov P, Kiritchenko S, et al. Semeval-2015 task 10: Sentiment analysis in twitter[C]//Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval. 2015.

[7] Zhang Y, Zhang H, Zhang M, et al. Do users rate or review?: boost phrase-level sentiment labeling with review-level sentiment classification[C]//Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014: 1027-1030.

[8] Tang D, Wei F, Qin B, et al. Building large-scale twitter-specific sentiment lexicon: A representation learning approach[C]// Proceedings of COLING. 2014: 172-182.

[9] Zhao Y, Qin B, Liu T, et al. Aspect-Object Alignment with Integer Linear Programming in Opinion Mining[J]. 2015.

[10] 徐琳宏，林鸿飞，潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.

[11] Jun Li and Maosong Sun, Experimental Study on Sentiment Classification of Chinese Review using Machine Learning

Techniques, in Proceding of IEEE NLPKE 2007.

[12] 赵妍妍, 秦兵, 刘挺. 大规模情感词典的构建及其在情感分类中的应用[C]//Proceedings of the 3rd National Conference of Social Media Processing (SMP2014), Beijing, China, 2014: 175-186.

[13] Labutov I, Lipson H. Re-embedding words[C]//Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics, 2013: 489-493.

[14] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for twitter sentiment classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014, 1: 1555-1565.

[15] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[16] Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 13-16.

[17] Büchner M H M P M, Stein B. Webis: An Ensemble for Twitter Sentiment Detection[C]//Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015: 582–589.

[18] 李泽魁, 赵妍妍, 秦兵, 等. 中文微博情感倾向分析特征工程[J]山西大学学报, 自然科学版, 2014, 37(4):570-579.