

# 基于中文bag-of-opinions方法的微博情感分析

王金刚<sup>1,2</sup> 于潇<sup>1</sup> 宋丹丹<sup>1,2,†</sup> 廖乐健<sup>1,2</sup> 邹维<sup>1</sup>

1. 北京理工大学计算机学院, 北京 100081; 2. 北京市海量语言信息处理与云计算应用工程技术研究中心, 北京 100081;

† 通讯作者, E-mail: sdd@bit.edu.cn

**摘要** 针对微博等新兴社交媒体的文本倾向性分析已经成为自然语言处理中的一个热点问题。基于依存文法提出中文 bag-of-opinions 模型表示微博文本, 将微博中的句子形式化地表示为情感评价单元的组合, 每个情感评价单元称为一个 opinion, 为每个 opinion 计算情感倾向值, 通过加权求和得到整个句子的情感倾向值, 从而实现对句子的观点识别和倾向性判断。此外, 为每个句子的情感倾向值赋予置信度, 根据置信度排序, 选择置信度高的句子作为标注语料, 解决微博文本标注语料不足的问题, 以便使用半监督或监督学习方法提升分类效果。在 NLP&CC 2012 微博情感分析评测语料上的实验结果也验证了该方法的有效性。

**关键词** 情感分析; 依存文法; 中文 bag-of-opinions; 微博

中图分类号 TN914

## A Chinese Bag-of-Opinions Method for Weibo Sentiment Analysis

WANG Jingang<sup>1,2</sup>, YU Xiao<sup>1</sup>, SONG Dandan<sup>1,2,†</sup>, LIAO Lejian<sup>1,2</sup>, ZOU Wei<sup>1</sup>

1. School of Computer Science, Beijing Institute of Technology, Beijing 100081; 2. Lab of High Volume Language Information Technology & Cloud Computing, Beijing, 100081;

†Corresponding Author, E-mail: sdd@bit.edu.cn

**Abstract** With the rapid growth of Weibo, sentiment analysis on it has become a hot topic in Natural Language Processing community. This paper proposes a Chinese bag-of-opinions model to represent Weibo sentences based on dependency grammar, in which every sentence is represent as a bag of opinions. We calculate a sentiment polarity score for every opinion, and then get a weighted summation sentiment evaluation for each sentence. A confidence value of a sentence's polarity score is also defined. With it, we can extract sentences with high confidence score as labeled data which can guide further analysis. Experiments conducted on the evaluation set for Chinese sentiment analysis at NLP&CC 2012 validate the effectiveness of our method.

**Key words** sentiment analysis; Chinese bag-of-opinions; ; dependency grammar; Weibo

微博是一种通过关注机制分享实时信息的广播式社交网络体系<sup>[1]</sup>。最早出现的微博应用是美国的 Twitter, 目前国内新浪、网易、腾讯等互联网公司都提供微博服务。微博具有信息更新速度快, 传播迅速, 扩散范围广等特点, 这使它容易对一些特定话题, 特别是一些社会敏感话题形成强烈的聚焦, 及时迅速地把握普通大众微博的情感信息成为一项有趣而有意义的工作。

NLP&CC 2012中文微博情感分析评测包括三个子任务:(1)观点句识别, 针对每条微博中的各个句子, 判断出该句是观点句还是非观点句(2)情感倾向性判断, 判断微博中每条观点句的情感倾向: 正面 (Positive), 负面 (Negative), 中性 (Neutral)(3)情感要素抽取, 找出微博中每条观点句作者的评价对象, 即情感对象<sup>[2]</sup>。我们的系统参加了前两项任务并提交了结果。

## 1 综述

### 1.1 语料预处理

本次评测数据来自腾讯微博。评测数据全集包括 20 个话题, 每个话题采集大约 1500 条微博, 共约 30000

条微博。数据采用 xml 格式，已经预先切分好句子。评测数据中包括大量的 xml 标记等格式化信息和与微博情感不相关的数据。本系统首先对语料进行了去噪处理，去除评论、转发以及空语句等无关信息。

## 1.2 依存语法

依存语法，是指通过分析语言单位内成分之间的依存关系揭示其句法结构，主张句子中核心动词是支配其他成分的中心成分，而动词本身却不受其他任何成分的支配，所有受支配成分都以某种依存关系从属于支配者。本系统使用哈尔滨工业大学信息检索研究中心开发的LTP(Language Technology Platform)<sup>[3]</sup>对微博语句进行分词、词性标注、词义消歧以及依存语法分析等。图 1 给出了微博“报道相当不靠谱”的依存关系树。其中，“靠谱”是根节点，其他节点以某种结构依存于它。ADV表示“状中”结构，副词“相当”和“不”修饰“靠谱”；SBV表示“主谓”结构，HED表示“靠谱”为句子核心。

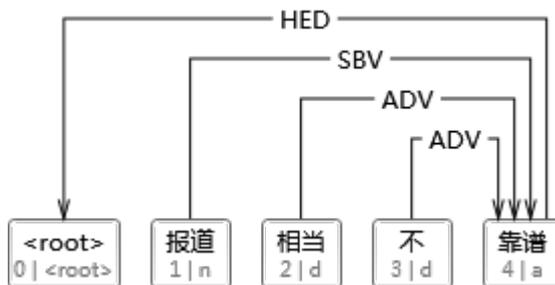


图 1 “报道相当不靠谱”的依存关系树

Fig.1 dependency tree of “报道相当不靠谱”

## 2 中文 bag-of-opinions 模型

### 2.1 中文 bag-of-opinions 模型

本节介绍基于依存语法的中文 bag-of-opinions 模型 (CBoO)。基于 CBoO 模型，可以把微博语句表示成 Opinion 的组合。定义 Opinion 为句子中最小的独立情感单元，每个 Opinion 由一个情感指示词(Indicator)、一系列修饰语 (Modifier, 包括否定词) 以及它们之间的依存关系 (Relationship) 组成。

每个 Opinion 被表示成若干个四元组：

$$Opinion = (Modifier, Indicator, Relationship, Offset) \quad (1)$$

或

$$Opinion = (Modifier_1, Modifier_2, Relationship, Offset) \quad (2)$$

其中，Modifier 为修饰词，Indicator 为情感指示词，Offset 为偏移（用来表示修饰词与情感指示词的距离），Relationship 为依存关系。

- Indicator（情感指示词）：句子中能够指示情感倾向的词。情感词典中的形容词、动词和名词可以作为情感指示词，例如：“美丽”（形容词），“鄙视”（动词），“混蛋”（名词）。
- Modifier（修饰词）：修饰情感指示词和其他的修饰词，一般是副词。在依存树中，修饰词是情感指示词或者其他修饰词的子结点。
- Relationship（依存关系）：表示情感指示词与修饰词或修饰词与修饰词之间的依存关系。例如：ADV 结构表示副词作状语修饰了情感指示词。
- Offset（偏移）：表示修饰词与情感指示词的距离，用一个有符号整数来表示。当修饰词出现在情感指示词的左侧时，offset 为负，否则为正。实际句子中，一个情感指示词可能被多个修饰词共同修饰，需要考虑修饰词先后顺序，所以考虑修饰词的位置。

图1给出的句子中，只有一个情感词“靠谱”，“靠谱”是情感指示词。另外三个结点（“相当”、“不”、“报道”）依存于它，是它的修饰词，“相当”和“不”以ADV关系依存于情感指示词。“报道相当不靠谱”可以表示成两个四元组：(相当, 靠谱, ADV, -2) 和 (不, 靠谱, ADV, -1)。其中，第二个元组的-1表示修饰词“不”在情感指示词的左侧，且距离是1。

## 2.2 情感倾向性计算

基于CBoO模型，系统可以计算出每条微博语句的情感倾向值，从而实现对微博语句的观点句识别和情感倾向分类。

### 2.2.1 情感指示词的倾向性值计算

情感指示词的情感倾向分为三种：正面，负面，中性。为每个情感指示词计算一个倾向值，不同的词拥有不同程度的情感倾向<sup>[4]</sup>。例如：“赞”和“不错”虽然都是正面倾向的词，但是他们的感情程度却不相同，“赞”的情感程度显然要比“不错”的情感程度强。但是由于目前缺乏公认的包含情感倾向程度的情感词典，系统中未考虑情感词的情感程度，这对最终的结果造成了一定影响。本系统中，正面情感词的倾向值统一设为+1，负面情感词的倾向值统一设为-1，中性词倾向值设为0，如公式(3)所示：

$$score_{indicator} = \begin{cases} +1, & \text{if } indicator \in \{positive\} \\ -1, & \text{if } indicator \in \{negative\} \\ 0, & \text{if } indicator \in \{neutral\} \end{cases} \quad (3)$$

### 2.2.2 修饰词的影响因子计算

系统将修饰词分为三类：强化修饰词 (Intensifier)，弱化修饰词(Reducer)和否定词(Negation)。强化修饰词会加强情感指示词或其他修饰词的强度；相反，弱化修饰词会减弱情感指示词或其他修饰词的强度。如“极其”是一个强化修饰词，“不够”是一个弱化修饰词。系统为每一个强化修饰词分配一个大于1的影响因子，为每一个弱化修饰词分配一个小于1的影响因子，如(4)式所示：

$$factor_{modifier} = \begin{cases} >1, & \text{if } modifier \in \{intensifier\} \\ <1, & \text{if } modifier \in \{reducer\} \end{cases} \quad (4)$$

否定词的处理比较特殊。当它出现在句子中时，情感倾向性取决于否定词和其他修饰词的顺序，不能简单地将情感倾向性反转。文献<sup>[6]、[7]</sup>中提到了英语中的这一语言现象，中文中同样存在这一现象，例如：“不是很漂亮”，如果简单将“很漂亮”的倾向性反转，将得到非常强的负面倾向性，实际上“不是很漂亮”是情感倾向较弱的正面短语。

当否定词依存于情感指示词时，其影响因子值为-1；当否定词依存于其他修饰词时，其影响因子取该修饰词影响因子的倒数。方程(5)表示当否定词出现时，影响因子的计算公式如下：

$$factor_{negation} = \begin{cases} -1, & \text{if } negation \text{ depends on an indicator} \\ 1 / factor_{modifier}, & \text{if } negation \text{ depends on a modifier} \end{cases} \quad (5)$$

情感指示词的倾向值通过修饰词的影响因子与情感指示词的情感倾向值相乘得到，如公式(6)所示：

$$score_{opinion} = factor_{modifier} * score_{indicator} \quad (6)$$

“不是很漂亮”中，score(漂亮)=+1，factor(很)=1.5，“不是”为否定修饰词，依存树如图2所示。

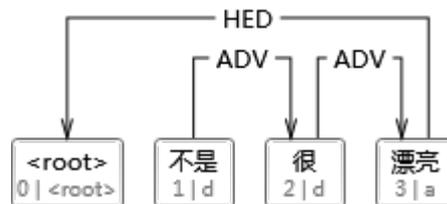


图2 “不是很漂亮”的依存关系树

Fig.2 dependency tree of “不是很漂亮”

系统从左向右计算该Opinion的整体得分。否定词“不是”依存于强化修饰词“很”，强化因子弱化为1/1.5，然后与“漂亮”的情感倾向值相乘，得到整体得分score(“不是很漂亮”) = 2/3。

### 2.2.3 句子的情感倾向值计算

当句子中所有Opinion的情感倾向值计算完成后，根据它们在依存树中的位置计算整个句子的情感倾

向值。在依存关系树中，离根节点近的 Opinion 的权重要比离根节点远的 Opinion 的权重大。本系统中，用情感指示词与根节点的距离表示该 Opinion 离根节点的距离  $pos(op)$ ，如 (7) 所示：

$$pos(op) = dis(indicator, root)。 \quad (7)$$

整个句子的情感倾向值计算公式如(8)所示：

$$score_{sentence} = \sum_{i=1}^N \frac{\sum_{i=1}^N pos(op_i)}{pos(op_i)} * score(op_i)。 \quad (8)$$

其中， $pos(op_i)$  为第*i*个情感指示词与根节点（root）的距离，N为句子中Opinion的数目， $score(op_i)$  为第*i*个情感指示词的情感倾向值。每条微博语句根据计算得到的情感倾向值分为正面，负面和中性。

#### 2.2.4 置信度计算

本次评测，我们提交了两组结果。其中，run2 使用了基于置信度的半监督聚类方法，为每个句子的情感倾向值定义一个介于 0~1 的置信度。该方法基于两个假设（1）具有单一情感倾向的句子得到的情感倾向值要比具有多种情感倾向的句子的情感倾向值置信度高（2）如果句子有多种情感倾向，正面情感倾向值和负面情感倾向值相差越大，句子的整体情感倾向值的置信度越高。置信度的计算公式如(9)：

$$confidence = \begin{cases} \frac{|score_{sentence}|}{|\sum_{i=1}^{N+} score(op_i)| + |\sum_{j=1}^{N-} score(op_j)|}, & \text{if } score_{sentence} \neq 0 \\ 1, & \text{if } score_{sentence} = 0 \end{cases}。 \quad (9)$$

其中， $N_+$ 为正面情感的个数， $N_-$ 为负面情感的个数。

### 3 实验与结果分析

#### 3.1 实验准备

实验数据集为NLP&CC2012中文微博情感倾向性分析评测数据集。本实验采用HowNet中文情感词典<sup>[7]</sup>和NTU情感词集<sup>[8]</sup>作为情感词典，并手工添加部分微博中常见的情感词如：“给力”，“脑残”等。实验中使用LTP<sup>[9]</sup>完成对原始语料的分词，词性标注，词义消歧以及依存句法分析等。

#### 3.2 实验方法

本实验使用三种方法来对微博语料进行主/客观分类和极性分类。（1）基准实验是基于情感词数目（Sentimental Word Count, SWC）直接分类的方法。任务 1 中，如果语句中存在情感词，则把语句分类为主观句，否则分类为客观句。任务 2 中，比较正面情感词和负面情感词的数目，选择数量较多的一方作为句子的整体情感倾向。（2）基于 CBoO 模型的直接分类。使用计算得到的情感倾向性值对句子的主客观性和句子的倾向性进行分类，这是一种无监督方法。（3）基于 CBoO 模型的聚类方法（CBoOC）。使用一元文法、标点、表情符作为上下文特征，使用情感词以及情感词数目作为情感特征，构建句子的特征向量。在任务 1 中，分别选择主观和客观中置信度高于特定阈值（实验中取 0.7）的若干句子作为初始簇，然后通过计算未分类向量与已知簇平均向量的欧几里得距离，对未分类的句子进行迭代聚类，直到每个簇中的句子数量不再变化为止，完成对所有微博句子的主客观分类。任务 2 中，选择三个类别中置信度高于某一阈值（实验中取 0.7）的若干句子作为初始簇，同样对未分类的句子进行迭代聚类，直到每个簇中的句子数量不再变化，完成对所有微博句子的极性分类。

#### 3.3 实验结果

使用准确率，召回率和F1值作为评测指标。考虑到评测数据中有20个话题，采用微平均(Micro\_Average)和宏平均(Macro\_Average)<sup>[10]</sup>两种方法来评价准确率、召回率和F1值。具体计算方法如下：

$$\text{微平均准确率: } MicroP = \frac{\sum_{j=1}^N correct(polarity = POS, NEG, NEU)}{\sum_{j=1}^N proposed(polarity = POS, NEG, NEU)} \quad (10)$$

$$\text{微平均召回率: } \text{MicroR} = \frac{\sum_{j=1}^N \text{correct}(\text{polarity} = \text{POS}, \text{NEG}, \text{NEU})}{\sum_{j=1}^N \text{annotated}(\text{polarity} = \text{POS}, \text{NEG}, \text{NEU})} \quad (11)$$

$$\text{微平均F1值: } \text{MicroF1} = \frac{\text{MicroP} * \text{MicroR} * 2}{\text{MicroP} + \text{MicroR}} \quad (12)$$

$$\text{宏平均准确率: } \text{MacroP} = \frac{1}{N} \sum_{i=1}^N \text{Precision}_i \quad (13)$$

$$\text{宏平均召回率: } \text{MacroR} = \frac{1}{N} \sum_{i=1}^N \text{Recall}_i \quad (14)$$

$$\text{宏平均F1值: } \text{MacroF1} = \frac{\text{MacroP} * \text{MacroR} * 2}{\text{MacroP} + \text{MacroR}} \quad (15)$$

任务1实验结果如表1:

表 1 句子的观点识别结果

Table 1 Subjectivity Classification Result

Method	Micro_Average			Macor_Average		
	Precision	Recall	F1	Precision	Recall	F1
SWC	52.0%	48.6%	50.2%	51.3%	47.2%	49.2%
CBoO	74.0%	55.7%	63.6%	73.3%	54.1%	61.8%
CBoOC	73.4%	52.8%	61.4%	72.4%	51.4%	59.9%

任务2实验结果如表2:

表 2 句子的倾向性判断结果

Table 1 Polarity Classification Result

Method	Micro_Average			Macor_Average		
	Precision	Recall	F1	Precision	Recall	F1
SWC	46.1%	48.6%	47.3%	45.5%	47.0%	46.2%
CBoO	72.4%	40.3%	51.8%	70.8%	38.7%	49.6%
CBoOC	71.8%	37.9%	49.6%	70.3%	36.5%	47.7%

### 3.4 讨论

任务1的评测结果显示, CBoO方法在准确率、召回率和F1值三项指标上都优于SWC方法, 特别是准确率一项, CBoO和CBoOC方法分别提高了22%、21.4%, 但是CBoO和CBoOC的召回率分别提高不足10%。任务2的评测结果也显示准确率提高较大(26.3%、25.7%), 但是召回率反而降低。考虑其可能的原因如下:

- (1) 实验中, 情感指示的词性限定为动词、名词和形容词, 但其他词性的词也可能作为情感指示词。另外, 中文中存在一词多词性现象, 例如:“激动”既可做动词, 也可作形容词, 如果词性标注模块错误地标注了词性, 将影响依存句法分析的结果, 从而对最终结果造成影响;
- (2) 如2.2.1节所述, 系统未对不同情感倾向程度的情感词做区分, 统一赋予固定的情感倾向值, 导致计算得到整句情感倾向值有一定误差。
- (3) 句子特征向量表示只使用了一元语法、表情符、情感词数目等特征, 并不能完整地表示文本深层的一些特征。

## 4 总结与展望

本文介绍了我们参加 NLP&CC 2012 中文微博情感分析评测的具体方法。从评测结果来看, 系统在两

项任务中的准确率较高，但是召回率比较低，同最好的结果相比还有一定的差距。今后还需要在方法上进一步改进，提升系统的召回率，进而提高系统的整体性能。

### 参考文献

- [1] <http://baike.baidu.com/view/1567099.htm>
- [2] 中国计算机学会中文信息技术专业委员会. “中文微博情感分析评测大纲”. 第一届CCF自然语言处理与中文计算会议 (NLP&CC) , 2012.
- [3] Che, W., Li, Z., Liu, T. Ltp: a chinese language technology platform. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations,2010, 13 - 16
- [4] Yao,T., Nie, Q. ,Li,J.,Li,L. Lou,D.,Chen,K. ,Fu,Y. An Opinion Mining System for Chinese Automobile Reviews. Frontiers of Information Processing. Tsinghua University Press,2005,268-289
- [5] Liu, J., Seneff, S.: Review sentiment scoring via a parse-and-paraphrase paradigm. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,2009, Volume 1 - Volume:161 - 169
- [6] Qu, L., Ifrim, G., Weikum, G.: The bag-of-opinions method for review rating prediction from sparse text patterns.Proceedings of the 23rd International Conference on Computational Linguistics, 2010, 913 - 921
- [7] <http://www.keenage.com>
- [8] <http://nlg18.csie.ntu.edu.tw:8080/opinion/>
- [9] <http://ir.hit.edu.cn/ltp/>
- [10] Zhao,S.,Zhang, Y.,LIU,T.,Chen,Y.,Huang,Y.,LI,S.A Feature Selection Method Based on Class Feature Domains for Text Categorization. Journal of Chinese Information Processing.2005, 23-29