基于 SVM 和 CRF 多特征组合的微博情感分析*

李婷婷，姬东鸿
(武汉大学 计算机学院，武汉 430000)

摘 要： 近年来，文本的情感分析一直都是自然语言处理领域所研究的热点问题；微博作为一种短文本，用词精炼而简洁，富含观点、倾向和态度。因此，识别微博的情感倾向具有重要的现实意义。提出一种基于 SVM 和 CRF 的情感分析方法，使用多种文本特征，包括词、词性、情感词、否定词、程度副词和特殊符号等，并选用不同的特征组合，通过多组实验使情感分析效果最优。实验显示，选用词性、情感词和否定词的特征组合时，SVM 模型的正确率达到 88.72%，选用情感词、否定词、程度副词和特殊符号的特征组合时，CRF 模型的正确率达到 90.44%。

关键词： 微博；情感分析；支持向量机；条件随机场
中图分类号： TP391.1 **文献标志码：** A **文章编号：** 1001-3695(2015)04-0978-04
doi:10.3969/j.issn.1001-3695.2015.04.004

Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features

LI Ting-ting, JI Dong-hong
(*School of Computer, Wuhan University, Wuhan* 430000, *China*)

Abstract: In recent years, the text sentiment analysis has always been a hot issue in the field of natural language processing. As a short text, micro-blog is featured of refined and concise, rich in views, tendencies and attitudes. Thus, the identification of emotional tendencies has important practical significance. This paper proposed a method of sentiment analysis based on SVM and CRF, used various features including word, speech, emotional word, negative word, adverb of degree and special symbols. They designed different combinations of features to make the effect optimal through multiple sets of experiments. The accuracy of SVM reached 88.72% using the combination of speech, sentiment word and negative word, while CRF attained 90.44% selecting the combination of sentiment word, negative word, adverb of degree and special symbols.
Key words: micro-blog; sentiment analysis; SVM; CRF

0 引言

微博是一种新兴的、开放的互联网社交服务，注册微博账号后，就可以通过关注机制分享实时信息。用户既可以作为观众阅览感兴趣的信息，又可以作为发布者发布信息。随着微博在中国网民中的日益火热，各种热词从微博中产生，并迅速蹿红网络，微博效应正在慢慢形成。据统计，截至 2013 年上半年，新浪微博的注册用户已达到 5.36 亿，微博正在逐渐渗透并影响中国网民的生活，成为中国网民在互联网上的主要活动之一。

情感分析又称为观点挖掘，即从文本中挖掘用户所要表达的观点及情感倾向。根据文本的不同粒度，情感分析任务可以分为篇章级[1]、句子级[2-4]和词语级[5]。随着注册用户的急剧增加，微博的影响力日益增大，国内外很多学者都已经积极参与到对微博的研究工作中[6-8]。

目前，国内关于中文微博情感分析的研究尚处于起步阶段，针对中文微博的语料库还很少。国外对于微博的情感分析已经做出一些尝试和探索，但将英文微博的情感分析方法具体应用到中文微博却存在一定的局限性。比如，虽然同样是 140 字的限制，中文所蕴涵的信息比英文更丰富；中文和英文的语法规则和语言习惯存在很大的不同，更强调上下文语境；英文不存在分词问题，但这对中文却是一个很大的挑战，尤其是面对微博这种口语化的短文本，其中存在大量未登录词，使得分词难度更大。

1 相关工作

随着微博的迅速发展，基于微博短文本的情感分析得到了越来越多国内外学者的关注，相关研究也陆续展开。本章将分别对情感分析的两种研究思路进行介绍，即基于情感知识的方法及基于特征分类的方法。

1.1 基于情感知识的方法

基于情感知识的方法需要建设情感词典或领域性情感词库，人工收集带有情感色彩的词语，并将其分为正向情感词和负向情感词，正向情感词包括“严谨”“美丽”“善良”等，负向情感词包括“糟糕”“阴沉”“邪恶”等。对于需要判定情感倾

收稿日期：2014-03-05；修回日期：2014-04-20 基金项目：国家自然科学基金重点项目(61133012)；国家自然科学基金面上项目(61173062)

作者简介：李婷婷(1990-)，女，湖北随州人，硕士，主要研究方向为自然语言处理(a393308100@163.com)；姬东鸿(1966-)，男，教授，博士，主要研究方向为自然语言处理、计算语言学.

向的文本,首先统计文本中正向情感词和负向情感词的个数,根据其差值判断情感倾向。这种方法虽然简单直观,但是在应用过程中存在很大局限性。首先,很难将所有的情感词完全收集,而且互联网络中不断有网络新词产生,现有词语的极性也可能随时间改变。此外,有些词语在不同的语境下可能表现出不同的情感极性,无法简单赋予其某种情感极性。

1.2　基于特征分类的方法

1) 有监督的机器学习方法

这类方法使用机器学习的模型,用已标注的训练数据训练出一个较好的模型,然后利用这个模型来预测文本的情感极性。机器学习模型包括支持向量机(support vector machine) 、朴素贝叶斯(naïve Bayes,NB) 、最大熵(max entropy,ME) 等。Pang 等人[9] 首先将机器学习的方法应用到文本的情感分析中,将情感分析看做一个二分类问题,包括正向和负向两类。他们首先对文本进行预处理,将对情感分析无用的信息过滤掉,然后从文本中抽取出多种特征,包括 unigram、bigram、词性特征、词的位置特征等,最后在 SVM、NB、ME 上分别进行实验。通过对比实验结果,发现选取 unigram 作为特征并使用 SVM 模型时效果最好。但是,Cui 等人[10] 后来通过实验证明,Pang 的实验结论具有一定局限性。当训练数据较少时,unigram 的效果比较好; 但随着训练数据的增多,unigram 的实验效果变差,n-gram($n>3$) 的实验效果不断优化,并超过 unigram 的实验效果。Kouloumis 等人[11] 的实验中显示,微博特有的表情和缩写常带有情感色彩,比情感词典的作用更大,但词性特征可能会成为情感分析的噪声。

2) 无监督的机器学习方法

Turney[12] 将无监督的机器学习方法应用到情感分析中,提出一种基于点互信息的方法,用以分析特定短语的情感极性,进而判断整篇评论的情感倾向。该方法首先选出两个词,正向词为 excellent,负向词为 poor,将其作为基本情感词。对于待分析的文本,先进行分词和词性标注,将其中带有形容词和副词的短语提取出来,然后使用 PMI 分别计算每个短语的正向关联度和负向关联度,最后根据文本中所有短语的正向关联度和负向关联度的差值来判断文本的情感倾向性。但是,无监督机器学习方法的正确率较低,实验效果依赖于基本情感词的选择以及评测语料的领域范围。

2　情感词典构建

2.1　预处理

微博中的#话题#、URL 和@ 用户等信息并不包含用户的观点,还可能成为下一步的分词和词性标注工作的噪声,对分词结果产生负面影响。因此在分词之前,要过滤掉微博中的#话题#、URL 和@ 用户等无用信息,然后再使用分词工具对微博进行分词和词性标注。本文使用的分词工具是 NLPIR2013,由中科院的张华平博士等人设计开发。

2.2　情感词表

在情感词典的构建中,本文主要参考台湾大学的情感词词典,过滤其中单字的情感词语和情感倾向不明显的词语,并人工挑选出一些情感词语以及网络新词进行补充,包括 2 953 个正向情感词和 4 096 个负向情感词。此外,将微博中常用的表情符号进行分类,包括 20 个正向表情符号和 24 个负向表情符号。所有的正向情感词和正向表情符号被赋值为 1,负向情感词和负向表情符号被赋值为 -1。

2.3　否定词表

否定词的出现往往会改变情感词的极性,收集了一些常见的否定词用于情感倾向的判定。表 1 列出了部分否定词。

表 1　部分否定词

不	非	无	甭	不要
不是	并非	毫无	切勿	不够
绝不	没有	从没	尚未	绝非

2.4　程度副词表

程度副词可以增强或减弱情感词的情感强度,本文也收集了一些常见的程度副词。按照程度副词的语气强度分别赋予强度值 0.5 到 2。表 2 列出了部分程度副词及其强度值。

表 2　部分程度副词及其强度

程度副词	强度值	程度副词	强度值
最,最为	2.0	比较,较为	1.0
特别,非常,极其,极端,很	1.5	有点,稍微,稍显	0.5

3　算法设计

本文研究的是对于主观信息的情感分析问题,即褒义和贬义的分类,不包括客观信息的判定。首先对微博进行预处理,然后使用 NLPIR2013 分词工具进行分词以及词性标注,抽取文本中的特征,分别在 SVM 模型和 CRF 模型上进行实验。

实验使用的是 COAE2014 中任务 4 所提供的语料,包含 2 174个训练语料和 5 000 个测试语料。表 3 统计了训练语料和测试语料中正向语料和负向语料的分布。

表 3　语料的统计结果

语料	正向语料个数	负向语料个数	总数
训练语料	993	1 181	2 174
测试语料	2 656	2 344	5 000

3.1　基于 SVM 的情感分析

SVM 模型是由 Vapnik 等人于 1995 年提出的,这种模型在解决小样本、非线性以及高位模式识别等问题中很有优势,并且能够应用到函数拟合等其他机器学习任务中。本文使用的工具包是 LIBSVM,由台湾大学林智仁教授等人设计开发。

3.1.1　特征选择

为了训练出一个较好的 SVM 模型,选取了五类文本特征,包括词性、情感词、否定词、程度副词及特殊符号。表 4 列出了 SVM 模型的所有特征类型及其含义。

表 4　SVM 模型的特征类型及其含义

特征类型	含义
词性	形容词、动词、名词、副词、介词、叹词的个数
情感词	正向情感词和负向情感词的个数、强度及情感得分
否定词	情感词之前是否出现否定词
程度副词	情感词之前是否出现程度副词
特殊符号	问号和感叹号的个数

3.1.2　特征抽取

对文本进行预处理、分词以及词性标注后,需要将表 4 中列举的五类特征抽取出来,然后使用 SVM 模型进行训练和预测。需要注意的是,由于 COAE2014 提供的训练语料中没有出现表情符号,而测试语料中多次出现表情符号,所以本文将正向表情符号作为正向情感词处理,将负向表情符号作为负向情

感词处理。否定词的出现可能改变情感词的情感极性,如果情感词之前出现否定词,则情感词的情感极性反向。比如,“这部电影不好看”,正向情感词“好看”前出现否定词“不”,所以“好看”的情感极性由正向变为负向。程度副词可以增强或减弱情感词的情感强度,本文收集了四类程度副词,分别赋值为0.5~2.0,如果情感词之前的一个词是情感副词,则情感词的情感强度相应改变。情感词典中正向情感词和负向情感词的初始值分别为1和-1。比如,“她长得特别漂亮”,正向情感词“漂亮”之前出现程度副词“特别”,“特别”的强度为1.5,所以“漂亮”的情感强度由1变成1.5。本文还将情感得分作为特征,情感得分是指文本中所有情感词的情感强度值之和,其中正向情感词的强度值为正值,负向情感词的强度值为负值。

3.1.3 特征组合设计

实验选取的文本特征包括词性、情感词、否定词、程度副词和特殊符号,其中否定词、程度副词和特殊符号单独作为模型的特征没有实际意义,需要与情感词搭配使用。为了找出最优特征组合,评估每种特征对SVM模型作用的大小,首先将词性和情感词作为特征进行两组实验,然后将词性和情感词作为特征组合进行实验,最后在词性和情感词的特征组合中分别加入否定词、程度副词和特殊符号特征。通过多组特征组合的实验,能够评估出不同特征对SVM模型的作用,并找出最优特征组合。

3.1.4 实验结果

实验使用正确率作为评估指标,表5列出了不同特征组合的实验结果。

表5 SVM模型中不同特征组合的实验结果

实验	特征组合	正面语料正确率/%	负面语料正确率/%	整体正确率/%
1	词性	56.40	68.43	62.04
2	情感词	88.14	82.59	85.54
3	词性+情感词	88.37	87.16	87.80
4	词性+情感词+否定词	87.95	89.59	88.72
5	词性+情感词+否定词+程度副词	87.84	88.87	88.32
6	词性+情感词+否定词+特殊符号	88.79	88.25	88.54

3.1.5 实验分析

实验结果显示,当使用词性、情感词和否定词的组合特征时效果最好,正确率达到88.72%。其中情感词的作用最大,使正确率提高了25.76%;其次是词性,使正确率提高了2.46%;否定词也对情感分析起到一定作用,使正确率提高了0.92%。但是,加入程度副词和特殊符号作为特征后,正确率略微降低,说明程度副词和特殊符号并不适合作为SVM模型的特征。

3.2 基于CRF的情感分析

CRF模型[13]是由Lafferty等人于2001年提出,结合了隐马尔可夫模型和最大熵模型的特点,在给定输入节点的条件下可以计算输出节点的条件概率,是一种无向图模型。CRF模型近年来被广泛应用于众多领域,包括自然语言处理、生物信息学等,在分词、词性标注以及命名实体识别等序列标注任务中均取得了很好的效果。本文将CRF模型应用到文本情感分析中,将情感倾向的判定看做一个标注任务,取得不错的效果。

3.2.1 特征选择

由于CRF一般用于序列标注任务中,而文本情感分析是要判断整个句子的情感倾向,并不是一个典型的标注任务。为了将情感分析转换成标注问题,本文将文本的极性对应到文本中每个词语的极性,通过标注每个词语的极性来判断文本的极性。换句话说,如果一条微博的极性为正向,则将微博中每个词语的极性都标注为正向;如果一条微博的极性为负向,则将微博中每个词语的极性都标注为负向。反之,如果一条微博中每个词语的极性都被预测成正向,则微博的极性就被判定为正向;如果一条微博中每个词语都被预测成负向,则微博的极性就被判定为负向。

除了表4所列举的五类特征外,CRF模型还使用了词本身的特征。由于SVM模型是典型的分类器,而CRF模型常用于序列标注问题,因此两种模型对特征的使用方法有所不同。表6列出了所有特征类型及其含义。

表6 CRF模型的特征类型及其含义

特征类型	值
词	词本身
词性	n,v,a,d,nr,…
情感词极性	1:正向,-1:负向,0:中性
否定词	1:是否定词,0:不是否定词
程度副词	0.5~2.0:程度副词的强度值,0:不是程度副词
特殊符号	2:问号,1:感叹号,0:不是问号或感叹号

3.2.2 特征模板

在CRF模型中,特征模板的选择对实验效果非常重要,本文使用的是开源工具包CRF++ 0.58。

这里简单介绍一下CRF++ 0.58所使用的特征模板格式:%x[row,col]。其中row和col分别表示相对的行偏移和列偏移,当前标记值的行偏移和列偏移均为0。举个例子,模板%x[0,0]表示当前行的词本身,模板%x[0,1]表示当前行的词性;此外,还可以使用特征的组合,如模板%x[0,0]/%x[0,1]表示当前行的词本身和词性的组合特征。

3.2.3 特征组合设计

实验选取的文本特征包括词、词性、情感词极性、否定词、程度副词和特殊符号。为了找出最优特征组合,评估每种特征对CRF模型作用的大小,首先将词作为特征进行实验,发现实验效果在正面语料和负面语料中很不平衡。依次加入词性和情感词极性特征后,实验的不平衡性得到改善,但效果依然不理想。笔者怀疑词和词性特征对CRF模型有干扰作用,所以单独使用情感词极性特征进行实验,实验的不平衡性消失且整体正确率得到提高,从而验证了词和词性特征对CRF模型的干扰作用。然后依次加入否定词、程度副词和特殊符号特征分别进行实验;最后将所有特征作为特征组合进行实验,再次验证词和词性特征对CRF模型的干扰作用。表7列出了本文使用的全部特征组合。

表7 CRF模型使用的特征组合

U00:%x[-2,0]	U09:%x[1,1]
U01:%x[-1,0]	U10:%x[0,2]
U02:%x[0,0]	U11:%x[-1,3]
U03:%x[1,0]	U12:%x[0,3]
U04:%x[2,0]	U13:%x[-1,3]/%x[0,2]
U05:%x[-1,0]/%x[0,0]	U14:%x[0,4]
U06:%x[0,0]/%x[1,0]	U15:%x[-1,4]/%x[0,2]
U07:%x[-1,1]	U16:%x[0,6]
U08:%x[0,1]	

3.2.4 实验结果

实验使用正确率作为评估指标,表8列出了不同特征组合的实验结果。

3.2.5　实验分析

实验结果显示,当使用情感极性、否定词、程度副词和特殊符号的组合特征时效果最好,正确率达到 90.44%。其中情感词极性的作用最大,仅使用情感词极性作为特征时,正确率就达到 88.40%;其次是否定词,使正确率提高了 1.84%;程度副词和特殊符号的作用较小,加入特征组合后使正确率提高了 0.2%。笔者还发现,传统方法中经常使用的词和词性特征并不适合作为 CRF 模型的特征,仅使用情感极性作为特征时正确率为 88.40%,加入词和词性特征后,正确率变为 87.02%,降低了 1.48%,而且使 CRF 模型对正面语料和负面语料的效果不平衡,前者高达 97.81%,后者仅为 74.78%。

表 8　CRF 模型中不同特征组合的实验结果

实验	选用特征	正面语料正确率/%	负面语料正确率/%	整体正确率/%
1	词	96.16	50.64	74.82
2	词 + 词性	95.78	53.11	75.78
3	词 + 词性 + 情感极性	97.81	74.78	87.02
4	情感极性	88.47	88.31	88.40
5	情感极性 + 否定词	90.47	89.97	90.24
6	情感极性 + 否定词 + 程度副词	90.40	90.36	90.38
7	情感极性 + 否定词 + 程度副词 + 特殊符号	90.51	90.36	90.44
8	所有特征	97.59	76.07	87.50

3.3　两种模型的对比

使用词性、情感词和否定词的组合特征时,SVM 模型的效果最好,正确率为 88.72%。使用情感极性、否定词、程度副词和特殊符号的组合特征时,CRF 模型的效果最好,正确率为 90.44%。将这两组实验进行对比,图 1 列出了这两组实验在正向语料、负向语料和全部语料的正确率。

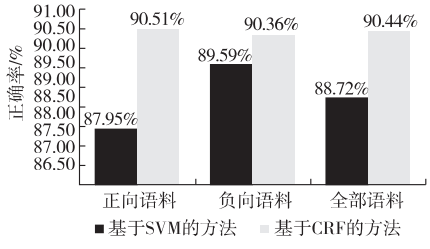

图 1　两种方法的正确率对比

从图 1 可以看出,基于 CRF 的方法在正向语料、负向语料和全部语料中均优于基于 SVM 的方法。这是因为 CRF 模型不仅使用节点自身的信息,还会使用上下文的特征和组合特征,使情感倾向的判定效果比 SVM 模型更好。通过对两种方法的最优特征组合进行对比发现,情感词和否定词特征对两种模型均有用;词性特征对 SVM 模型有用,对 CRF 模型有干扰作用,程度副词和特殊符号特征对 CRF 模型有用,对 SVM 模型有干扰作用。表 9 列出了实验所选取的特征对两种模型的作用,1 表示该特征对模型有用,-1 表示该特征对模型有干扰作用。

表 9　各特征对模型的作用

特征	对 SVM 模型的作用	对 CRF 模型的作用
词性	1	-1
情感词极性	1	1
否定词	1	1
程度副词	-1	1
特殊符号	-1	1

4　结束语

本文提出一种基于 SVM 和 CRF 多特征组合的情感分析方法,在 COAE2014 所提供的语料上取得了很好的效果,SVM 模型的正确率达到 88.72%,CRF 模型的正确率达到 90.44%。通过实验发现,除了传统的词性和情感词极性等特征外,否定词、程度副词、特殊符号等特征也起到一定作用,但是,不同的机器学习模型适合不同的特征组合。程度副词和特殊符号特征在 SVM 模型中不适用,会使实验的正确率下降。情感分析经常用到的词和词性特征在 CRF 模型中不适用,会使实验效果在正面语料和负面语料中产生不平衡的现象。

笔者将继续深入研究,进一步提升实验效果,将进行的工作如下:

a) 实验证明情感词典对实验效果至关重要,目前的情感词典还不够完整,将继续丰富情感词典。

b) 由于网络新词的不断产生,人工完善情感词典仍然无法满足需求。因此,新情感词的自动发现和极性判定将是以后重点研究的问题之一。

c) 笔者还将加入更多特征训练机器学习模型,如句法特征及语义特征等,寻找更好的特征组合。

参考文献:

[1] BALAHUR A, STEINBERGER R, KABADJOV M, *et al*. Sentiment analysis in the news [J]. Infrared Physics and Technology, 2014, 65: 94-102.

[2] JIANG Long, YU Mo, ZHOU Ming, *et al*. Target-dependent twitter sentiment classification [C] // Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.

[3] 王金刚, 于潇, 宋丹丹, 等. 基于中文 bag-of-opinions 方法的微博情感分析 [C] //NLP&CC. 2012.

[4] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis and opinion mining [C] //Proc of International Conference on Language Resources and Evaluation. 2010.

[5] TABOADA M, BROOKE J, TOFILOSKI M, *et al*. Lexicon-based methods for sentiment analysis [J]. Computational Linguistics, 2011, 37(2): 267-307.

[6] 谢丽星, 周明, 孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取 [J]. 中文信息学报, 2012, 26(1): 73-83.

[7] LUCIANO B, FENG Jun-lan. Robust sentiment detection on twitter from biased and noisy data [C] // Proc of the 23rd International Conference on Computational Linguistics. 2010.

[8] 李寿山, 黄居仁. 基于 Stacking 组合分类方法的中文情感分类研究 [J]. 中文信息学报, 2010, 24(5): 56-61.

[9] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [C] //Proc of Conference on Empirical Methods in Natural Language Processing. 2002: 79-86.

[10] CUI Hang, MITTAL V, DATAR M. Comparative experiments on sentiment classification for online product reviews [C] //Proc of the 21st National Conference on Artificial Intelligence. 2006: 1265-1270.

[11] KOULOUMIS E, WILSON T, MOORE J. Twitter sentiment analysis: the good the bad and the OMG! [C] // Proc of the 5th International AAAI Conference on Weblogs and Social Media. 2011: 538-541.

[12] TURNEY P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C] // Proc of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 417-424.

[13] SUTTON C, MCCALLUM A. An introduction to conditional random fields for relational learning [M] // Introduction to Statistical Relational Learning. Cambridge: MIT Press, 2006.