基于 SVM 的微博文本情感倾向性识别

韩忠明,张慧,解筱梦

(单位北京工商大学计算机与信息工程学院　北京 100048)

摘要：

本文针对微博数据进行观点句判别及情感倾向性分类进行深入研究。本文以 HowNet 情感分析用词表作为基本词典，过滤其中的单字词语，并进行网络情感词补充等，构成一个情感词典。使用基于支持向量机（SVM）分类方法，优化情感特征项的选取，训练样本，对被测数据进行较为准确的预测。在第一届 CCF 自然语言处理与中文计算会议情感分析评测中，本文针对观点句判别及观点句情感倾向判定取得了较好的效果。

关键词：支持向量机；情感词典；特征选择；

Effective Sentiment Classification Method Based on SVM for Microblogging Texts

Han Zhongming, Zhang Hui, Xie Xiaomeng

(School of Computer Science and Information Engineering, Beijing Technology and Business University Beijing 100048)

Abstract： In this paper, we focus on sentiment classification method for Microblogging texts, Based on HowNet emotional lexicons, an emotional lexicon is constructed after unrelated words are filtered. We propose a optimization feature selection method and thus classify texts using SVM classifier. In emotional evaluation competition of the first CCF Conference on Natural Language Processing & Chinese Computing, proposed method demonstrated good performance in terms of precision and recall.

Keywords: SVM, Emotional lexicons; Feature section;

0 引言

本文的目的在于提出一种优化的基于 SVM 方法判别微博观点及判别微博情感倾向性的方法。本文借鉴一般文本分类方法，进行算法的改进。构建微博情感词典，对测试数据集进行预处理，噪声处理，优化选取特征项，通过样本模型化处理，计算得出测试数据是否为观点句及观点句的情感倾向分类。

文本倾向性分析技术不仅可以应用于微博分析，对购物反馈，产品评论、网络舆情检测及垃圾消息过滤等领域也有着广泛的应用，通过判别文本的情感倾向可以指导用户购买某种产品、监控网络舆情等，现有的中文倾向性分析主要研究定位于对句子或者段落等进行判别。针对微博式短文本，产品评论，电影评论，网络即时消息，论坛等的情感倾向性分析研究较少，本文以微博为例进行分析。

1 相关工作

文本情感倾向分析目的在于，判别自然语言中表达的情感倾向。许多文本情感分析主要针对中长文本，对于微博这样的短文本处理方法较少。国内外对于文本情感倾向性的研究大体上分为两大类：基于语义的文本情感倾向性研究和基于机器学习的文本情感倾向性研究。

（1）基于语义的文本情感倾向性研究。

2011 年，何凤英[1]以 HowNet 情感词语集为基准，构建中文基础情感词典，计算并标注情感词的极性。利用词典及程度，副词词典结合情感词极性值计算文档句子情感值来获取文本的情感倾向性.考虑了语言风格及结构，但是对于微博短文本，表达情感的句式结构非常少，甚至没有，主要的一些词就可以表达情感。2011 年 Yue Lu[2]等人提出一种学习不同来源数据，结合上下文自动构建情感词典的算法。Yue Lu 等对于情感词典的构建进行了新的扩展及改进。

（2）基于机器学习的文本情感倾向性研究

2011 年，Dmitriy[4]等人提出基于 N-gram 情感分类方法。使用数据中的长短短语作为特征值对文本进行情感分类。此方法使一些具有情感意义的组合词，发挥他们情感倾向的比重意义。但是，对于微博短文本，几个字组成的情感表述，效果不明显。2010 年，咎红英[5]等人将机器学习中的经典分类方法与规则方法相结合，用以分析新闻语音文本的情感倾向，判断其强弱。通过 SVM 分类器来研究特征选择方法及特征权重计算方法的组合对实验结果的影响。本文基于 SVM 分类方法，针对微博数据特性进行特征选择及权重计算，进而判断情感倾向分类。

2 任务分析

本文实验数据来源于第一届 CCF 自然语言处理与中文计算会议中文微博情感分析测评，测评对象是面向中文微博的情感分析核心技术，包括观点句识别、情感倾向性分析和情感要素抽取。本论文参与任务一，任务二的测评工作。

2.1 任务一观点句识别

针对每条微博中的各个句子，本任务要求判断出该句是观点句还是非观点句。如表 1 所示。有 3 条微博，每条微博有一个句子。对每个句子进行观点句标注。显然 weibo1，weibo2 中的 sentence1 为观点句，weibo3 中的 sentence1 不是观点句。

表 1 观点句识别样例

< weibo id="1"><sentence id="1" opinionated="Y">#洗碗工留剩菜被开除# /奋斗洗碗工人做的对，全国人民要支持她!! </sentence> </weibo>
<weibo id="2"><sentence id="1" opinionated="Y">#洗碗工留剩菜被开除#这是什么烂酒店:不分清红皂白。 </sentence> </weibo>
-<weibo id="3"><sentence id="1" opinionated="N">一个国家经济出了问题没关系可要是思想出了问题就不好办了! </sentence> </weibo>

2.2 任务二观点句情感倾向标注

本任务要求判别观点句的情感倾向。观点句的情感倾向可以分为正面（POS），负面（NEG）和其他（OTHER）。如表 2 所示。对 weibo 中的 sentence 进行情感判别。如 webo1 中的 sentence1 判定为正面（POS），如 webo2 中的 sentence1 判定为负面（NEG）。

表 2 观点句情感倾向标注样例

<weibo id="1"><sentence id="1" polarity="POS" opinionated="Y"> #洗碗工留剩菜被开除# /奋斗洗碗工人做的对，全国人民要支持她!! </sentence> </weibo>
<weibo id="2"> <sentence id="1" polarity="NEG" opinionated="Y"> #洗碗工留剩菜被开除#这是什么烂酒店:不分清红皂白。 </sentence> </weibo>

3 微博观点句及情感识别方法

本论文的任务是识别情感观点句，并且判别情感倾向。本文认为，带有情感倾向的句子所表达的看法就是观点句。因此，判别是否为观点句，就要计算句子中的情感成分。本文方法基于 SVM 机器学习方法，使样本包含非观点句，观点正面情感句，观点负面情感句。对样本进行模型化，并预测测试数据结果。可得出非观点句，观点句，并对观点句进行情感标注。本文方法简称为 DSVM 方法。

DSVM 方法对传统的 SVM 文本分类方法，针对微博数据长度短，文本表述不规范，噪音多等特点，进行了改进。主要从特征词的选取，结合预处理后的数据，通过特定类别观点词和情感词制定的情感词典，经过噪音处理筛选后，计算词语的权值，然后选取特征项。并根据特征项对样本文本进行向量化处理，进而模型化处理。使测试数据通过预处理，噪声处理后，通过特征项进行向量化，并根据样本模型对测试数据进行预测。最终得出结果。这种方法大大的提高了微博数据的观点句识别及情感倾向性判别的准确性，并降低了算法的复杂度问题。

本文中情感词典不同于传统情感词典，对于词典结构进行了结构简化处理，提高情感词遍历及检索的速度。去除情感词典中对情感词倾向性标注，依赖 SVM 的样本模型化对文本进行倾向性判别。

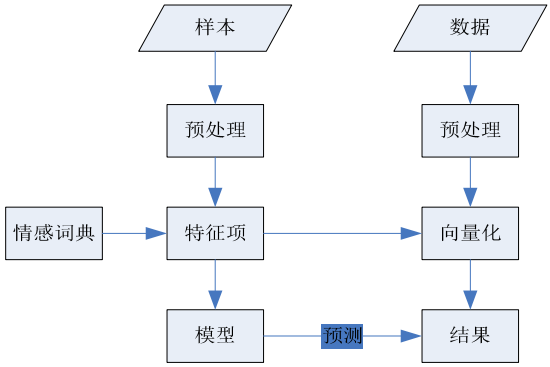

图 1 方法流程图

主要流程为图 1 所示。样本选取，提取特征词，训练样本为模型，可进行测试数据的预测处理。对样本训练成模型，需要进行特征词的选取。本论文选取 K 个特征词。特征词为情感词典中的情感词汇。并将测试数据经过预处理之后，根据特征词进行向量化处理，最后依据 SVM 进行预测，进而得到结果。

3.1 情感词典构建

本论文采用的情感分析方法，使用具有高权重的情感词作为特征值，使用 SVM 方法训练样本。本文认为，具有情感色彩的文本可以作为观点。因此识别观点可以通过识别情感进行判别。使用情感词典中的情感词作为特征项。情感词典以 HowNet 作为基本词表，并融入当今热点网络流行情感词汇构成情感词典。例如:“凶残”，“善良”，“坑爹”等。具有明显情感色彩的词语。剔除传统词典中对于情感词的标注倾向性的结构，只保留情感词。

3.2 样本数据

样本数据全部来源新浪微博。其中没有情感观点倾向数据条数 200 条，标记为 NONE，具有负面情感观点倾向条数为 728，标记为 NEG，具有正面情感倾向条数为 541，标记为 POS。共计样本条数为 1496。其中部分来源于，会议提供的样本数据。样本数据内容涵盖实物，事物，及人物方面的数据。

3.3 数据预处理

微博数据，句式短，结构少，用词复杂，随意性强，表达关键及情感主要集中在一些词，及部分带有情感的语句中。因此，只要发现利用这些核心的词语就可以判

别观点及情感倾向。语料库数据来自微博，含有许多噪音，处理过滤掉无价值的数据，可以减轻算法计算量，缩短计算时间，还可以保证准确性。本文预处理方法如表 3 所示。

表 3 预处理方法

```
*输入：一条微博内容
*输出：观点词，情感词序列
1.data=Pretreatment(weibodata) #处理字符编码，过滤标点符号
2.swords= ICTCLAS(data) #使用中科院分词软件进行分词标注
3.sentences=filter（swords）#过滤停用词，单字词
4.count=0     #用于记录整条情感观点词数
5.FOR s IN sentences:
6.      Sen=Dic(s)
#记录词表里情感词，词典中不包含的词汇，略掉
7.      calculate=Calculate(halfs)     #统计词相关值
8.      count=count+1
9.END FOR
```

3.4 特征选择及数据向量化

特征的提取，来源于样本数据。针对样本数据首先进行预处理，进行分词处理，词性标注处理。本论文方法是用是 ICTCLAS 分析系统，并根据以往大量实验人工校准词典。对每条样本数据进行过滤处理，只留下名词，动词，形容词。如上节预处理所示。

(1) 计算每个词 w 的 A,B,C,D 值。

A:包含w并且所在样本数据属于 POS 的文本个数。

B:包含w并且所在样本数据属于 NEG 的文本个数。

C:不包含w的样本数据，并且属于 POS 的文本个数。

D:不包含w的样本数据，并且属于 NEG 的文本个数。

(2) 对每个w计算它的x^2估计（CHI）：

x^2计算的是特征w与类别 C 之间的依赖关系。如果w与 C 之间相互独立，那么文本特征 w 的x^2估计值为零。对于类别 C，文本特征 w 的 x^2 估计定义如下：

$$x^2 = \frac{\mathrm{N} \times (\mathrm{A} \times \mathrm{D} - \mathrm{C} \times \mathrm{B})}{(\mathrm{A}+\mathrm{C}) \times (\mathrm{B}+\mathrm{D}) \times (\mathrm{A}+\mathrm{B}) \times (\mathrm{C}+\mathrm{D})}$$

N为样本数据

(3) x^2 排序选取前K个。则K个 w 作为特征项。

数据向量化，对样本和测试数据依据选出的特征词进行向量化处理。计算每个数据文本中的 K 个特征项的权重。

(1) 统计每个特征项在该文本数据中的个数m。记 TF

(2) 计算每个w的权重值。

$$\mathrm{TF} \times \mathrm{IDF} = \mathrm{m} \times \lg \frac{\mathrm{N}}{\mathrm{A}+\mathrm{B}}$$

(3) 将样本及测试数据向量化存储。

3.5 建模及测试数据预测

本文 DSVM 方法对样本向量化结果进行模型化处理，模型出非观点句，观点正面句，观点负面句的特征结构，形成样本模型。对测试数据进行预处理，噪声处理，利用特征项将测试数据进行向量化。使用模型，对向量化后的测试数据进行预测。得出非观点句，观点情感正面句，观点情感负面句。

4 实验结果与分析

本实验数据来源于“2012 年 CCF 自然语言处理与中文计算会议”提供的测评数据集。数据集包含 20 个不同话题的数据。数据集共 31675 条数据，平均每个话题 1584 条数据。测评数据以 XML 格式存储，总数据大小为 5.64MB。

本论文对测试数据分别进行三种方法的实验，即方法 1：基于 HowNet 的计算方法，方法 2：基于词典方法，方法 3：本实验方法。使用词典均为同一情感词典。

方法一

基于 HowNet[2] 的方法（简称：HN 方法）。在 HowNet 中每个词语有多种表达含义，则词语有多个义项,例如：词语“好”，可表达“易于”，“健壮”，“喜欢”等义项。每个义项又由多个义原组成，那么词语的语义相似度计算实际上是由义原的相似度计算得到的。测试数据情感值不等于 0 为观点句（等于 0 为非观点句），大于 0 为正面，小于 0 为负面。

方法二

基于情感词典的方法（简称:ED 方法）。使用基于 HowNet 词表为基础词表，进行情感词典的补充，添加程度词，副词，转折递进词并标注权重。对数据只留下词典中包含的词语并对词语进行权重标记并计算测试数据的情感值。测试数据情感值不等于 0 为观点句，大于 0 为正面，小于 0 为负面。

4.1 实验结果评估指标

在对方法的有效性进行评估时，本文使用的评估指标为：准确率（Precision），召回率（Recall），F 值（F-measure）。微平均以整个数据集为一个评价单元，计算整体的评价指标；宏平均以每个话题为一个评价单元，计算参评系统在该话题中的评价指标，最后计算所有话题上各指标的平均值。

4.2 实验结果分析

1.观点句识别结果

图 2 和图 3 分别给出了观点句识别中微平均和宏平均下的 3 个指标值。

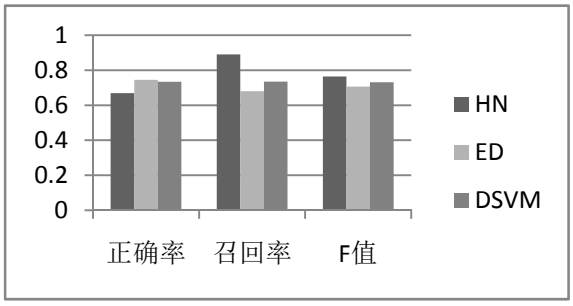

图 2 宏平均观点句识别结果

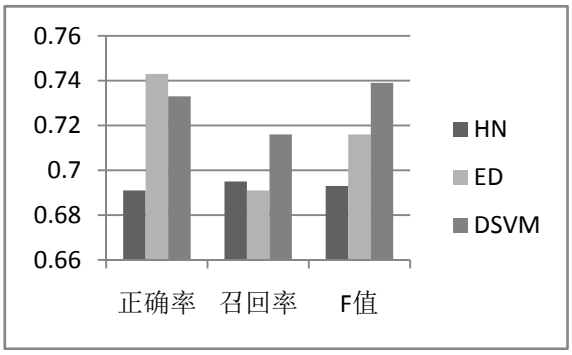

图 3 微平均观点句识别结果

如图 2 所示，宏平均中 HN 方法（基于 HowNet 的方法）显示效果较好，但由于很低的正确率，过高的召回率，致使 F 值效果好。而在图 3 微平均中，本实验（DSVM）效果，无论是从准确率，还是召回率，较方法 HN，ED 都要好。通过实验得出，本实验方法对观点句的识别有较好的效果。本实验针对微博数据不同于博文，及一般文本数据。将自然语言问题，转化为数学概率问题。避免许多句式结构及词语表示的歧义问题。所以本文方法 DSVM 在观点句的识别中有较好的效果。

2.情感倾向性判别结果

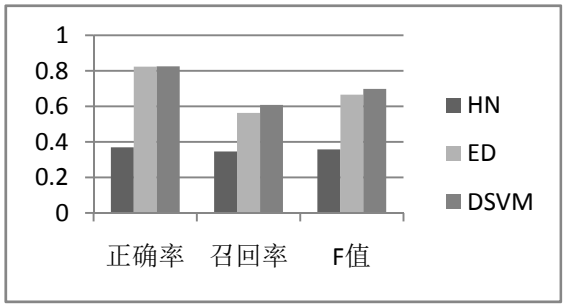

图 4 宏平均情感倾向性判别结果

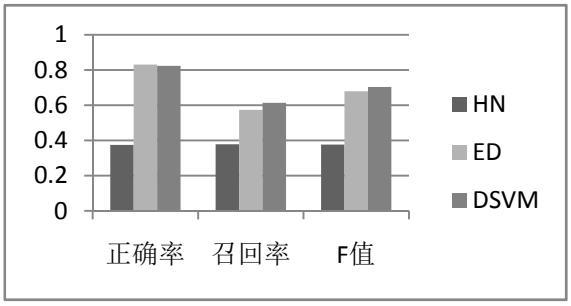

图 5 微平均情感倾向性判别结果

由图 4，图 5 情感倾向性所示，无论宏平均，还是微平均，DSVM 的效果较前两个方法都有较高的正确率，召回率，精度值。

实验表明，基于知网的 HN 方法看起来考虑的很全面，综合考虑了情感词、否定词、程度副词、疑问句、感叹句、总结句和转折递进复句类型，并且利用同义词词典进行相似度计算，来确定一个句子的倾向性。但实际结果的正负是由正负情感词个数决定的，不能正确理解句子的语义。准确率很差，并且复杂程度高，时间冗余。

基于词典的 ED 方法，是对基于知网方法的精简，考虑了句子的语义信息。但是过于依赖词典，对于新兴情感词汇识别相比较差，需要人工更新词典，效果较次。

本文基于 SVM 的 DSVM 方法，针对微博数据的表述更加随意性，许多结构及表述方式不同于常态表述，并极具个性化表述的特点。进行预处理，特征词的选取，向量化处理。在测试中分类效果较好，虽然依赖情感词典选取特征项，但是融入 SVM 对于句子结构，及词语概率的考虑。可以弥补一些未收录到情感词典的词为发挥的作用。

5 结论及展望

本文以 HowNet 的情感词典为基础，构建情感词典，基于 SVM 机器学习方法，对特征项的选取做出改进，提出了一种计算短文本情感倾向性的方法。实验部分，将本文提出的方法与基于 HowNet 的典型方法、基于词典的方法进行比较，验证了本文提出方法的有效性。本文提出的下一步的工作是完善情感词典，改进此方法，自动识别新兴词汇，补充情感词典。加强样本学习，继续优化特征项的选取，从而更好地识别情感观点句及情感倾向分类。

参考文献：

[1]何凤英，基于语义理解的中文博文倾向性分析，计算机应用，31（8）2011

[2] Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach

[3]杨超，冯时，王大玲，杨楠，于戈，基于情感词典扩展技术的网络舆情倾向性分析，小型微型计算机系统，2010，31（4）

[4]Sentiment Classification Based on Supervised Latent n-gram Analysis

[5] 昝红英，郭明等，新闻报道文本的情感倾向性研究，计算机工程，36(15) 2010

[6] 丁琼． 基于向量空间模型的文本自动分类系统的研究与实现［D］．上海：同济大学，2007．

[7] C. Buckley and et al., “The smart/empire tipster ir system," in Proceedings of TIPSTER Phase IIIWorkshop, 1999.

[8]丁建立, 慈祥, 黄剑雄. 网络评论倾向性分析. 计算机应用, 2010, 30(11):2937-2940

[9]刘群,李素建.基于《知网》的词汇语义相似度计算. ，第三届汉语词汇语义学研讨会,台北,2002 年 5 月