基于多分类器集成的中文微博情感分析

潘艳茜　罗珞　杨慧　张楠　李武军†

上海交通大学计算机系，上海，200240

† 通讯作者，E-mail: liwujun@cs.sjtu.edu.cn

摘要　针对 2012 年 CCF 自然语言处理与中文计算会议（NLP&CC 2012）面向中文微博的情感分析中的任务 1（观点句识别）和任务 2（情感倾向性分析），使用新浪微博中与社会新闻有关的微博作为训练数据，以 Support Vector Machine 和 Naïve Bayes 分类器为核心，提出了一种分类器集成的方法，将微博中的句子分为观点句和非观点句并判断观点句的情感倾向性。实验结果表明，集成分类器可以达到比单一分类器更好的效果，在评测数据上测试任务 1 和 任务 2 的 F 值可分别达到 0.767 和 0.759，处于较好水平。

关键词　微博情感分析；观点句识别；情感倾向性分析；分类器集成

A Classifier Ensemble Approach for Chinese Microblogs Sentiment Analysis

PAN Yanxi, LUO Luo, YANG Hui, ZHANG Nan, LI Wujun†

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240

†Corresponding Author, E-mail: liwujun@cs.sjtu.edu.cn

Abstract　This paper focuses on Task 1 (opinionated sentence identification) and Task 2 (polarity determination) in NLP&CC 2012. By using microblogs concerning social news from sina as training data, based on Support Vector Machine and Naïve Bayes, a classifier ensemble approach is proposed to classify microblogs into subjective class and objective class and determine the polarity of opinionated sentence. Experimental results show that ensemble classifier outperforms single classifier. The F scores of Task 1 and Task 2 are 0.767 and 0.759 when test on evaluation data.

Key words　sentiment analysis of microblogs; opinionated sentence identification; polarity determination; classifier ensemble

随着互联网的发展，越来越多用户通过网络平台表达自己观点，从而产生了大量的主观性文本数据。这些数据中蕴含的大量情感信息，其具有非常大的潜在价值，在社会舆情分析、有害信息过滤、产品推荐等诸多领域有着广阔的发展前景。然而这些数据的主观性情感分类无法通过传统的基于关键词和自动索引信息获取，而人工浏览大量文本又十分低效。近年来，针对文本的情感分析是一个研究热点，相关技术已在电子产品、影视娱乐和新闻等多个领域得到应用[1]。

微博信息是一种在社交网络上通过关注机制分享的简短实时信息，其内容具有时效性，主题包罗万象，且拥有海量数据。基于微博的情感分析技术可以在各个领域提供有用信息。但与传统的情感分析不同，微博由于其内容过于简短（如新浪微博不超过 140 字），用户发言含各种噪声（如错别字，非正式用语等）等因素，对其进行情感分析相比传统的在产品评论等领域的相关工作要困难得多。近年来国际上有关英文 Twitter[1]的情感分析研究较为热门。但基于中文微博的相关工作相对较少，中文微博中经常使用反讽等方式表达情感，这也使得中文微博的情感分析更为困难。

本文介绍了我们的中文微博情感分析系统，该系统中我们采用了 Stanford Word

[1] https://twitter.com/

Segmenter[1]对中文微博进行分词预处理，通过 Stanford Parser[2]得到文本的语义信息，并使用机器学习的方法对微博进行情感分析。第一部分介绍了相关领域的研究现状；第二部分描述了本系统所采用的观点句识别方法；第三部分介绍系统的情感倾向性判断方法；第四部分介绍相关实验；第五部分为全文总结。

1 相关研究

情感分析主要有两个方面工作，一是从大量文本中抽取带有感情色彩的主观性文本，即主观/非主观的分类问题；二是抽取带有情感色彩的主观性文本的有价值的情感信息，如正面或负面的情感分类、评价对象抽取等。在目前的研究中，情感分析被应用于各种不同粒度的文本。之前大量的工作是针对文档级别的文本进行分类，比如判断一段电影评论是正面还是负面[2]。还有一些工作针对于句子级别的文本分类[3]。特别的，一些研究专门分析关于产品各个特征的观点[4]。此外，还有一些工作涉及词和短语级别的情感分析，比如建立一个高质量的情感词典，这个词典可能是主题无关的[5]，也可能是主题相关的[6,7]。

关于情感分析的研究方法，主要有两种：基于词典的语义方法和机器学习的方法。语义方法通常通过计算候选词和通用情感词典中的基准词的语义距离，判断候选词的情感倾向。例如，Lu 等将不同来源的信息结合起来构成一个统一的最优框架，这些信息包括通用情感词典中该词的极性，整个文档的情感分数，WordNet 中的同义词、反义词信息，以及一些语法规则（比如两个用“和”连接的词的极性也更相近）[8]。另一方面，机器学习的方法首先需要人工标记情感语料库，然后训练出一个模型来学习出不同类别的特征，从而预测目标文本属于哪一类别。用于训练的特征通常包括 unigrams，bigrams，词性和词的位置等等。分类算法主要有监督学习和半监督学习两种。常用的监督学习的方法包括 SVM（Support Vector Machine），NB（Naïve Bayes），最大熵（Maximum Entropy）[2]和 KNN（K-nearest neighbor）。一些半监督学习的方法应用了 bootstrap 策略，比如自我训练和联合训练[9]。

传统的情感分析处理的都是一些比较规范的文本，如产品评论或博客等，然而，社交网络的情感分析处理的是内容较短且语法不规范的文本。判断微博的主客观和情感倾向性比传统的情感分析要困难的多。目前，国外关于 Twitter 的研究日益增多，其中包括对每条 tweet 的情感分析，对于一个话题的情感分析[10]，以及用户级别的情感分析[11]。在 2011 年，Jiang 等提出了一种与情感对象有关，上下文相关的方法来判断 tweet 的情感[12]。

2 观点句识别

2.1 建立语料库

实验中所用到的训练集由两部分组成，中国计算机学会（CCF）所提供的来自腾讯微博的样例数据[3]和我们自己抓取的新浪微博的数据，主题均与社会新闻相关，包括菲军舰恶意撞击、疯狂的大葱、官员财产公示等 26 个话题，共 3552 条微博，6508 个句子。训练集为人工标注，由两个人分别独立标记，结果不同的再由第三个人进行裁决，以尽量避免由于个人理解不同造成的误差。

2.2 数据预处理

由于句子的主客观性与其带有的 hashtag 没有必然联系，为不影响分类效果，我们首先去掉了句子中的 hashtag，仅保留句子本身的内容。此外，在分词之前，我们对句子中的特殊符号进行了正规化处理，如将所有全角符号转化为半角，将英文标点替换为中文标点，将连续的数字替换为<NUM>，将所有不规范的省略号替换为<ETC>，还将短链接替换为<SHORT_URL>，以避免分词带来的误差。

对所有句子进行正规化后，我们用 Stanford Word Segmenter 对句子行分词，利用 Stanford Parser 得到词性及词与词之间的依赖关系。

1 http://nlp.stanford.edu/software/segmenter.shtml
2 http://nlp.stanford.edu/software/lex-parser.shtml
3 http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html

2.3 基于 SVM 的分类器

SVM 模型将数据表示为向量 $x^{(i)}$，即空间中的点，在这个空间中建立一个可以将不同类别的向量分开的超平面（法向量表示为 $\mathbf{w}$），并且使得两类之间的间隔最大。其目标函数如(1)所示：

$$min_{r,w,b}\frac{1}{2}\|w\|^2, \tag{1}$$

$$s.t.\ y^{(i)}(w^T x^{(i)}+b)\geq 1,\ i=1,\ldots,m,$$

其中 $y^{(i)}\in\{1,-1\}$，为数据点 $x^{(i)}$ 的类别。通过解一个对偶问题，可以得到最优解为(2)式:

$$w=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} \tag{2}$$

α_i 是拉格朗日算子。只有位于边界上的点的 $\alpha_i>0$，称为支持向量，其余点的 $\alpha_i=0$。故只有支持向量决定了超平面的函数。分类器通过计算测试数据落于超平面哪一侧，来判断数据的类别。

试验中，我们采用了 Chih-Jen Lin 的 LIBLINEAR（2007）工具包[1]进行观点句非观点句分类，应用线性核函数，在 bag-of-words 的基础上尝试了以下不同的特征集合：

1. unigram 的 binary 表示，若该单词出现，则特征向量的相应维的值为 1，否则为 0。
2. unigram 的 tf 值表示，$w_{t,d}=\begin{cases}1+\log tf_{t,d}, & if\ \ tf_{t,d}>0\\ 0, & otherwise\end{cases}$，其中 tf 为该单词在句子中的词频。特征向量的每一维为该单词的 w 值，并按向量进行正规化，使得每个句子对应的特征向量成为单位向量。
3. unigram 的 tf-idf 值，$w'_{t,d}=(1+\log tf_{t,d})\times\log(N/df_t)$，其中 N 为总的句子数，df 为包含该单词的句子数。特征向量的每一维为该单词的 w' 值，并进行正规化。
4. bigram 的 tf-idf 值，为了减少分词带来的误差，并考虑词之间的相互联系，此特征在 unigram 基础上，加入了两个相邻词组合而成的 bigram 的信息。为防止特征向量过于稀疏，我们采用了用互信息进行特征选择，仅保留与观点句和非观点句互信息高的词进行降维并提高准确率。采用的互信息的公式为(3)：
$$(x,y)=\log(\frac{p(x\mid y)}{p(x)})=\log\frac{p(xy)}{p(x)p(y)} \tag{3}$$
其中 $p(x)$ 为词 x 在训练数据中出现的概率，$p(y)$ 为属于类别 y 的句子的概率。
5. 特征集合参考了[13]中的方法，选取了 URL、不同类别词个数、特殊句式等 9 个与区分观点句和非观点句密切相关的特征，如表 1 所示。其中主张词、连词、代词和程度副词的特征来自 HowNet 词典[2]。评价词和评价对象根据[14]中的方法生成。大致过程如下，首先选取“挺好”、“不错”、“荒唐”等 52 个常用形容词以及与微博 hashtag 相关的“政府”、“海军”、“中国”等 66 个名词作为种子，构成最初的评价词和评价对象集合，然后通过两个集合中词和集合之外词之间的语法依赖关系对集合进行扩展，直到集合中无法加入新的词为止，最终得到我们所需要的所有评价词和评价对象。

[1] http://www.csie.ntu.edu.tw/~cjlin/liblinear/
[2] http://www.keenage.com/

表 1 观点句识别特征集

Table 1 Feature set for opinionated sentence identification

编号	特征描述	样例
1	是否含 url	含有 http
2	评价词个数	低劣、惊艳
3	评价对象个数	安徽官、奥迪车
4	主张词个数	觉得、认为
5	连词个数	况且、然而
6	代词个数	我们、他
7	程度副词个数	非常、十分
8	是否为特殊句式	疑问句、感叹句
9	是否为引用	含有引号、书名号

2.4 基于 NB 的分类器

Naïve Bayes 算法将一个句子 s 分为概率最大的那一类 $c^*_{NB}(s) = \underset{c}{argmax}\, P(c \mid s)$。利用 Bayes 公式计算 $P(c \mid s)$:

$$P(c \mid s) = \frac{P(s \mid c)P(c)}{P(s)} \tag{4}$$

通过假设在给定类别的条件下，句子中的每个词 x_i 相互条件独立，NB 将 $P(s \mid c)$ 分解为：

$$P(s \mid c) = \prod_{x_i \in s} P(x_i \mid c) \tag{5}$$

其中 $P(x_i \mid c)$ 为每个词在某一类中出现的频率，经过 Laplace smoothing 得到。

2.5 结合 SVM 和 NB 的集成分类

使用上述基于 SVM 和 Naive Bayes 两种分类器，对每个句子可分别得到一个分类结果（$L \in \{1,-1\}$，1 和-1 分别表示观点句和非观点句）和属于该类别的置信度（$cf \in [0,1]$）。考虑单个分类器带来的误差，我们采用了两种方法将两个分类器得到的结果集成，一种是将两种分类器的分类结果和置信度作为更高一层集成分类器的特征进行再分类，另一种是直接对置信度加权平均得到最终的分类结果。

2.5.1 两层分类器

将训练集划分为 Basic 和 Ensemble 两个集合，在 Basic 集合上训练得到 SVM 和 NB 基本分类器。然后使用基本分类器在 Ensemble 集合上进行分类，将基本分类器得到的结果和置信度以及句子的实际类别作为样本，构建训练集，作为集成分类器的训练数据。我们采用了神经网络[1]和 SVM 两种方式训练集成分类器。

2.5.2 加权平均的集成分类器

本系统采用公式(6)得到集成分类器：

[1] http://neuroph.sourceforge.net/

$$P(Y) = \alpha \times cf_{SVM}(Y) + \frac{(1-\alpha) \times L_{NB} \times |cf_{NB}(Y) - cf_{NB}(N)|}{\max(cf_{NB}(Y), cf_{NB}(N))} \quad (6)$$

其中α为调节 SVM 和 NB 分类器结果所占权重的参数。若最终得到的$P(Y)$大于特定阈值，则认为句子是观点句，否则是非观点句。

3 情感倾向性分析

为了减小观点句识别时带来的误差，本系统直接将句子分为正面，负面和无观点三类，而没有在观点句识别的结果上再分类。并且在倾向性分类中，我们舍弃了 OTHER（中性及其它无法明确归为正面或者负面的）这一类，一是因为这一类数据非常少，训练数据中仅占观点句的 2%（50/2497），句子总数的 0.77%（50/6508）；二是因为这一类别的界定比较模糊。

首先，我们仍然分别采用基于 SVM 和基于 Naive Bayes 的分类器，将句子分为正面，负面和无观点三类。基于 SVM 的分类器原理如 2.3 节所述，选取 unigram 的 tf-idf 值为特征。基于 Naive Bayes 的分类器与 2.4 节所述类似。

然后，将 SVM 和 Naive Bayes 的分类结果采用加权平均公式(7)，(8)，(9)集成：

$$P(N) = \alpha_1 \times cf_{SVM}(N) + (1-\alpha_1) \times P_{NB}(N) \quad (7)$$

$$P(NEG) = (\alpha_2 \times cf_{SVM}(NEG) + (1-\alpha_2) \times P_{NB}(NEG)) \times \mu_{NEG} \quad (8)$$

$$P(POS) = (\alpha_3 \times cf_{SVM}(POS) + (1-\alpha_3) \times P_{NB}(POS)) \times \mu_{POS} \quad (9)$$

其中，$cf(N)$，$cf(NEG)$，$cf(POS)$分别为句子属于非观点句、负面和正面观点句的置信度。α_1，α_2，α_3为控制 SVM 和 NB 分类结果所占权重的参数，μ_{NEG}，μ_{POS}为调节三个类别置信度权重的参数。最终得到的 *P* 最大的那一类别为当前句子的分类结果。

4 实验结果

综合考虑各种方法在训练集上的性能，对于两个任务，我们最终都选取特征为 unigram 的 tf-idf 值，基于 SVM 分类器的分类结果为第一组提交结果，把将 SVM 和 NB 进行加权平均的集成分类器所得结果作为第二组提交结果。在评测数据上所用参数均为在训练集上交叉验证 F 值取得最大时的参数。其中任务 1 中集成分类所用的参数为：$\alpha = 0.8$，$threshold = 0.1$。任务 2 中集成分类所用参数为：$\alpha_1 = 0.4$，$\alpha_2 = 0.4$，$\alpha_3 = 0.6$，$\mu_{POS} = 2.1$，$\mu_{NEG} = 1.8$。

本系统在任务 1 取得的评测结果如表 2 所示。

表 2 任务 1 评测结果
Table 2 Task 1 evaluation results

结果编号	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
34	0.805	0.588	0.680	0.807	0.581	0.671
35	0.745	0.789	0.767	0.748	0.782	0.760

其中集成分类器的 F 值处于评测的较好水平，为第 6 名。基于 SVM 分类器的正确率处于评测的较好水平，为第 5 名。

本系统在任务 2 取得的评测结果如表 3 所示。

表 3 任务 2 评测结果
Table 3 Task 2 evaluation results

结果编号	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
34	0.867	0.510	0.642	0.872	0.507	0.636
35	0.861	0.679	0.759	0.866	0.676	0.757

其中集成分类器的召回率和 F 值均处于评测的较好水平，分别为第 5 名和第 3 名。基于 SVM 分类器的正确率处于评测的较好水平，为第 5 名。

从实验结果看，集成分类器的召回率和 F 值均比 SVM 单一分类器的值高，说明集成分类器确实结合了两个基本分类器的长处。SVM 分类器的正确率最高，但召回率不足。情感倾向性分析的性能不如观点句识别，这是由于错误地将非观点句分成了观点句从而影响正负极性的分类造成的。

5 总结

本文介绍了本小组参加 NLP &CCF 2012 中文微博情感分析评测的基本情况。我们首先从新浪和腾讯微博获取所需微博信息，通过人工标注构建训练数据集和测试数据集。然后将微博句子正规化，采用 The Stanford Natural Language Processing Group 的开源工具对句子进行分词、词性标注等处理，将数据转化为分类器的特征。系统分别使用 LIBLINEAR 实现的 SVM 分类器，我们自己实现的 Naive Bayes 分类器以及这两种分类器集成的方法对微博句子进行观点句识别。实验发现，基于 SVM 的分类器使用 unigram 的 tf-idf 值作为特征时分类的性能最好，仅凭借评价词个数等特征来识别观点句性能较差。将 SVM 和 NB 两个分类器的结果加权平均得到的集成分类器比单一分类器效果要好，有效地提高了 F 值。

参考文献

[1] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of ACL 2002.

[2] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In EMNLP, 2002.

[3] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In CICLing 2005, LNCS 3406, pages 486–497, 2005.

[4] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In WSDM’08, 2008.

[5] S. Mohammad, C. Dunne, and B. Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In EMNLP, pages 599–608, 2009.

[6] Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In ACL, pages 590–598, 2009.

[7] V. Jijkoun, M. d. Rijke, andW.Weerkamp. Generating focused topic-specific sentiment lexicons. In ACL, pages 585–594, 2010.

[8] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: An optimization approach. In WWW, 2011.

[9] X.Wan. Co-training for cross-lingual sentiment classification. In ACL, 2009.

[10] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In CIKM’11, 2011.

[11] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis

incorporating social networks. In KDD'11, 2011.

[12] L. Jiang, M. Yu, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In ACL, 2011.

[13] 徐睿峰，王亚伟，徐军，张玥，郑海清，桂林，叶璐. 基于多知识源融合和多分类器表决的中文观点分析. 第三届中文倾向性分析评测会议（COAE）济南 2011.

[14] Guang Qiu, Bing Liu, Jiajun Bu and Chun Che. Expanding Domain Sentiment Lexicon through Double Propagation. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)