COAE2015 中文微博文本情感分析*

李淼[1]，职启予[2]，李秋池[1]

[1]清华大学电子工程系，北京，100084

[2]北京邮电大学信息与通信工程学院，北京，100876

E-mail: miao-li10@mails.tsinghua.edu.cn

摘　要：本文主要对 COAE2015 评测中的任务二提交的算法进行描述，并结合评测结果做了分析与总结。针对于判断微博倾向性的任务，本文提出了三种方法，一种是基于情感词典的极性判断方法，一种是基于统计学习设计分类器的方法，这两种方法中我们都针对于限定资源和非限定资源提出了不同的分类器设计方法，最后一种是将二者融合的一个尝试。COAE2015 的评测结果表明我们针对微博情感分析的方法是有效的。

关键词：微博；情感词典；统计学习；融合

Chinese Weibo Sentiment Analysis for COAE2015

Miao Li, Qiyu Zhi, Qiuchi Li,

Multimedia Signal and Intelligent Information Processing Lab, Tsinghua University, Beijing, 100084

School of Information and Communication Engineering, BUPT, Beijing, 100876

E-mail: miao-li10@mails.tsinghua.edu.cn

Abstract: This paper aims to describe the algorithm proposed for the task 2 in COAE 2015, and make a conclusion according to the evaluation result. Three methods are proposed especially for Chinese weibo sentiment classification. The first is a polarity classification method based on sentiment lexicon, the second is to design a classifier from a statistical model. Two different classifiers are designed towards restricted and unrestricted resource in these two methods. The tentative fusion system, combining the first two system, is also submitted. The evaluation results demonstrate the effectiveness of these methods.

Keywords: weibo; sentiment lexicon; statistical learning; fusion

1　引言

情感分析是自然语言处理研究中一个很重要的领域。具有巨大的研究价值。目前，情感分析已经被应用在舆情分析、产品评价、股票评论、广告投放等各个领域中。

微博不仅可以即时通信，同时具有社会化媒体的特点。随着互联网的日益流行，微博显示出了它巨大的影响力。它有以下几个特点，第一，传播大量信息。据马茜[1]统计，截止 2014 年，新浪微博注册用户达 2 亿多，日均活跃账号数达 1 亿，每天用微博发送的信息超过 250,000,000 条。第二，消息源头。现如今大量娱乐、体育甚至政治事件都先从微博曝出。第三，与国际联系紧密。微博上我们可以关注名人，名企动态。除此以外，情感分析的技术还有助于文本摘要[2]、问答系统[3]等研究工作。

现在关于微博的情感分析研究现状主要是以下几种方法，基于词典的情感分析是用已有资源，如 WordNet 等构建情感词典， 然后去看文本中包含正向情感词和负向情感词的个数，然后判断句子的正负关系。Taboada 在[4]提出一种基于词典的方法 Semantic Orientation

*单击此处输入基金或资助机构的名称（项目编号），或删去此行

CALculator (SO-CAL) 去提取文本中的情感。SO-CAL 使用了一种标注极性和强度的情感词典，证明了这个模型的可行性并解释了构造词典的过程。Kanayama 在[5]为主题领域的情感分析构造了无监督的词典模型，结合上下文的连贯性构造“极性词原子”，然后调整阈值去估计句子的情感极性。

基于机器学习的方法是指先标注训练语料和测试语料，使用支撑向量机（SVM）、最大熵、KNN 等分类器进行情感分类。Mullen 在[6]中提出了一种基于 SVM 的情感分析方法，利用 unigram 特征以及统计词性最后在影评的语料上得到了不错的结果。另外 pang 在[7]中用情感分析来给文本分类，在影评数据上尝试了集中机器学习的方法，有朴素贝叶斯、最大熵和支撑向量机。

其他还有结合了主题模型方法的做法[8],另外 Prabowo[9]尝试了一种结合了规则分类、监督学习和机器学习的方法，在 Myspace 评论和产品评论的数据上取得了不错的效果。

本文综合使用了现在主流的基于词典的方法、基于机器学习的方法，另外还针对 COAE 评测尝试融合了两种方法。之后的文章在第二节会详细描述评测的任务，第三节介绍对微博语料预处理的步骤，第四节介绍基于词典的方法，第五节介绍基于机器学习的方法，第六节介绍系统融合的框架，第七节和第八节针对实验和评测结果对系统进行分析。

2 任务描述

本任务是给定已经进行了断句的微博，要求参赛系统自动分析微博中每个带观点句子的倾向性，即褒义、贬义、褒义贬义混合。所以每条句子有四种可能的情感倾向：褒义、贬义、褒义贬义混合以及无情感。评测只评价前三种的识别情况，微博观点的极性中 1 代表褒义，-1 代表贬义，0 代表既有褒义又有贬义。

3 预处理步骤

我们先去除了微博中一些罕见符号，比如不能用 utf-8 编码的特殊符号。这一步处理过之后只剩汉字英文数字以及常见符号。

然后对特殊符号做了处理，把 URL 替换成了“ URL ”，@替换成了“ ATsomeone ”,数字替换成了“ NUM ”，其他特殊符号替换成了“ NO ”。

接着，我们为了处理规整化，将所有全角符号替换成了半角符号，英文替换成了中文，同时进行了简繁体的转换，用的是 linux 下 opencc 的工具。

此外，我们进行了重要符号的统计，有 hashtag 的#，表情符号的[]，以及句子的逗号、句号、叹号、引号。

最后在分词的步骤，我们把上面替换的特殊字符加到了我们的词典，不至于分词系统把它分开，然后去停止词的工作也在分词时候完成。

我们最后对一条微博预处理的结果是 terms[],termclass[]两个列表，前一个储存分的所有词，后面储存对应的词性。

4 基于情感词典的方法

4.1 词典介绍

我们先使用 NTUSD, Hownet, CEWO, 中文负情感词典等融合了一个大的词典，融合方法是投票判断每个词的情感倾向，将每个词标为+1、-1 以及 0 情感。然后我们使用限定资源的情感词汇本体以及限定资源中的正面、负面情感词语以及正面、负面评价词语融合为同样的一个将每个词标为+1、-1 以及 0 情感的词典，融合方法是如果各个词典中对于一个

词的情感判断不同，那就放到之前融合的大词典中判断其情感倾向，最后只提取标注为+1和-1 的词语，最后的情感词典组成如下：

表 1 融合词典组成

极性	1	-1
数量	12714	12182

我们同样利用限定资源构造了副词词典，并且手工标注了不同的强度值，其组成如下：

表 2 副词词典组成

强度	2	1.5	1.2	1.1	0.8
数量	69	42	37	29	12
举例	极度、绝对	分外、格外	更加、较为	蛮、稍稍	不怎么

在 Unrestrict 的情况下，我们直接使用了之前融合的大词典，然后再融合了限定资源的词典，如果有大词典未出现的词就直接收录，这样融合后的情感词典组成如下：

表 3 大词典组成

极性	+1	-1
数量	27087	26005

另外除了使用和 Restrict 情况中相同的副词词典之外，我们还构造了包含 23 个否定词的否定词典和转折词典，否定词典具体内容见下表：

表 4 否定词典组成

不 不可 不能 不得 不要 不用 不够 不再 不必 没有 非 并非 无法 无 并未 未必 别 莫 难以 不宜 绝不 决不 毫无

表 5 转折词典组成

但是 但 然而 然 而 可是 可 不过 只是 就是 但是

4.2 分类器设计

对一条微博语料，先划分字句计算每个子句的情感值，我们用 e 表示各种组合下情感值得计算，p 代表情感词强度，包含 1 和-1，a 代表程度副词的程度值，那么各种组合计算情感值得方法如下表：

表 6 词组合情感值计算

序号	类型	情感值计算
1	只含有情感词	e=p
2	含有情感词和副词	e=a*p
3	含有情感词和否定词	E=-1*p

可以总结，每个分句中计算情感倾向的方法是：

$$E_j = \sum_{i=1}^{N_p+N_n} e_i$$

其中 e_i 是对每个情感词的组合的情感值计算，然后对整个分句中所有的情感词叠加得到整个分句的情感值。我们将阈值设为 0.9，意思是如果 E_j 的值大于 0.9，那么这个分句被判为 1，就是正情感，如果小于-0.9，那就是-1，即负情感，其他情况判为 2，即无情感。

统计整句微博所有分句的情感值：

$$E = \sum_{j=1}^{N} E_j$$

其中 E_j 代表微博子句的情感值，N 代表预料中句子数目，那么每个分句的有 1、-1 和 2 的情感值，这些情感值组成了整句微博语料的 list，最后根据这个 list 的组成来分类，判定方法如下：

$$E = \begin{cases} 0 & if\ 1\&-1\ in\ list \\ 2 & else\ if\ 1\&-1\ not\ in\ list \\ -1 & else\ if\ 1\&2\ not\ in\ list\ or\ -1\ in\ list \\ 1 & else\ if\ -1\&2\ not\ in\ list\ or\ 1\ in\ list \end{cases}$$

4.3 加入规则

我们同时在 unrestrict 中考虑加入了比较句和转折词的两个规则。关于比较句规则，我们参考引用了简单五种规则如下：

*不如 n+*v*　　*相比 v+*+n*　*与 p+*+相比 v*　　*不及 n+英文**　　*不及 n+中文**

其中 n 是名词，v 是动词，p 是介词，*表示任意词汇，在这五种规则出现在微博语料中时，我们默认语料中有正有负，即判为 0。

我们单独考虑了转折词存在的情况，当语句中出现转折词，我们单独用转折词来分割我们的微博语句，即分为转折词前和后的两句分句来分别判断分句的情感倾向。

5 基于机器学习的方法

5.1 特征选择

支撑向量机是普遍使用的一种训练统计模型分类器的方法，其中选取特征又是其中重要的一环。在特征选择部分，由于训练数据较少，所以选用的特征是 unigram 和情感词典的 unigram，采用逻辑回归分类器，用默认参数。

5.2 分类器设计

针对评测具体任务，训练数据选用 NLPCC2012,13,14 的数据。数据组成如下：

表 7 NLPCC 数据组成

情感	0	1	-1
数量	28533	8075	7622

本文分别训练了两个模型：一个是用正情感例子和中性情感的例子，一个是用负情感例子和中性情感的例子。保证正负例相同，每一次都是从中性例子中随机抽句子保证和正（或者负）情感例子相同。

这样得到了两个分类器，分别可以计算出句子属于正情感和负情感的概率。最后的分类策略是，如果属于正、负情感的概率均小于阈值 1，则判别为 2；否则，如果正、负情感的概率之差小于阈值 2，则判别为 0；否则，正负概率值那个大判别为哪个。

6 系统融合框架

在融合部分，对于非限定资源（unrestricted），SVM 部分的方法就是上述，融合的方法是，如果句子中检测出了连词或者比较句，则选择基于情感词典的方法；如果情感词典方法的判别是 0，则最终的结果就是 0，其他一律按照 SVM 方法输出。

对于限定资源，SVM 的方法中仅仅用了 unigram。没有用情感词典的 unigram，调优的最优参数依然是 0.4 和 0.05。对于融合的方法，也是按照非限定资源一样的策略。

7 实验以及结果

7.1 实验设置

我们在测试数据中选取了 500 句微博，手工标注情感值为 0、1、-1 以及 2。其中 0 表示有正有负，2 表示无情感。最后确定在这 500 句上调整以及确定参数。

在词典的方法里，判断正负情感的阈值设置为 0.9。另外同样经过在测试数据（随机选择的 500 句微博标注情感）上调参数，调得最后 SVM 方法里参数阈值 1 为 0.4，参数阈值 2 为 0.05。

7.2 实验结果

本文在任务二的测试集上进行试验，共计微博篇章 50000 篇，已将篇章切割成句子，共计 133201 句，最后的三种方法的评测结果如下：

表 8 限定资源 Restrict 评测结果

	Value(Lexicon)	Value(SVM)	Value(merged)	Medium	Best
Pos_F1	0.7581	0.6850	0.6591	0.6115	0.8526
Neg_F1	0.3625	0.6192	0.4308	0.5117	0.7191
Mix_F1	0.2000	0.1019	0.1786	0.1330	0.3892
Micro_F1	0.5892	0.6207	0.5210	0.5273	0.7705
Macro_F1	0.4888	0.4691	0.4611	0.4714	0.6543

表 9 非限定资源 Unrestrict 评测结果

	Value(Lexicon)	Value(SVM)	Value(fusion)	Medium	Best
Pos_F1	0.7762	0.6707	0.6591	0.6975	0.8742
Neg_F1	0.2328	0.6061	0.4308	0.5800	0.7868
Mix_F1	0.1864	0.2862	0.1786	0.1617	0.4233
Micro_F1	0.5946	0.6066	0.5210	0.6057	0.8113
Macro_F1	0.4801	0.5332	0.4611	4947	0.6945

如上图，给出了评测最终的结果以及各个评价指标的数值情况，在限定资源的情况下，提出的的词典和 SVM 的方法在 19 个最终结果排名第 5 和第 7，非限定资源也排到了中游的位置。但是融合的方法在两个结果中表现都在下游的排名位置。

关于词典的方法，Neg_F1 的表现尤其不好，总结是因为词典方法中分类器设置的问题，因为 0 是有正情感负情感所以造成了分类的不便，另外因为都是微博短句，可能会造成情感词的不足，以至于词典的方法并不能充分利用情感词的判断基准。

关于 SVM 的方法，由于训练集和测试集的不统一，以及开发集的规模小，导致训练的模型并没有完全将微博情感分类，尤其是 0 的情况，既有正情感负情感。训练集和测试集的不统一表现在 NLPCC 只有无情感的句子，并没有标注既有正情感负情感的句子。开发集的规模小是因为我们只标注了 500 条微博来调整参数。

关于融合的方法，因为我们在开发集上测试，发现词典的方法在负情感以及判断 0 的方面都不及 SVM 的方法，所以我们只判断有规则的句子，其他不含规则的句子全按照 SVM 的方法判断，评测结果表明这样是不可行的，下一步的工作会集中在更好的融合两种方法。

8 总结

本文针对 COAE2015 的评测采用了三种方法，分别是基于情感词典的方法，基于机器学习的方法以及一种融合的方法，前两种方法在最后的评测结果中取得了较好的效果，然后在一些方面两种方法都有需要改进的地方，下一步的重点可以是两个系统针对 COAE 评测的融合。

参 考 文 献

[1] 马茜. 新浪微博大 V 的传播分析. [硕士学位论文]. 郑州大学. 2014

[2] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 43-50.

[3] Wilson T, Hoffmann P, Somasundaran S, et al. OpinionFinder: A system for subjectivity analysis[C]//Proceedings of hlt/emnlp on interactive demonstrations. Association for Computational Linguistics, 2005: 34-35.

[4] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational linguistics, 2011, 37(2): 267-307.

[5] Melville P, Gryc W, Lawrence R D. Sentiment analysis of blogs by combining lexical knowledge with text classification[C]//Proceedings of the 15th ACM SIGKDD international

conference on Knowledge discovery and data mining. ACM, 2009: 1275-1284.

[6] Mullen T, Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources[C]//EMNLP. 2004, 4: 412-418.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[8] Lin C, He Y. Joint sentiment/topic model for sentiment analysis[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 375-384.

[9] Prabowo R, Thelwall M. Sentiment analysis: A combined approach[J]. Journal of Informetrics, 2009, 3(2): 143-157.