第3章

Sentiment Analysis: Mining Opinions, Sentiments, and Emotions

文档级情感分类

从本章开始，我们将讨论情感分析中的主要研究问题和当前所提出的最好算法。文档级情感分类(document sentiment classification)或者文档级别的情感分析(document-level sentiment analysis)可能是情感分析领域(尤其是早期)最为广泛研究的任务了(Pang and Lee, 2008; Liu, 2010)。这一任务的目标是将一篇给定观点的文档(如产品评论)根据所持观点为正面或负面进行分类。正面或负面观点又称情感的倾向性或极性。这个任务即文档级别的分析，因为它将一篇文档看作一个整体，且并不研究文档中具体的实体或属性，也不研究针对这些实体的情感倾向。尽管这一任务在情感分析领域被广泛研究，但是这一任务本身的局限使得在实际应用中更加需要细粒度的情感分析，即基于属性的情感分析(Hu and Liu, 2004)(第5、6章)

文档级情感分类是最简单的情感分析任务，因为它把情感分类当作传统的文本分类问题，只是类别变成了情感的倾向或者极性。因此，任何一种监督学习算法都可以直接应用到这个问题上来。在大多数情况下，其所使用的特征与传统文本分类使用的特征相同。由于问题定义简单，且与文本分类等价，因此这一任务是情感分类其他衍生任务的基础，而这些衍生任务也来源于传统文本分类任务，例如跨文档和跨领域的情感分类。

为确保这个任务有实际意义，已有的关于文档级情感分类的文献都约定如下假设(Liu, 2010)。

假设3.1：文档级情感分类假设观点文档 d(如一篇产品评论)表达的观点仅针对一个单独实体 e，且只包含一个观点持有者 h 的观点。

因此，严格说来，文档级情感分类只能用于一种专门类型的观点文档。我们在任务定义中明确指出了这个假设。

定义3.1(文档级情感分类)：给定针对一个实体的观点文档 d，判断观点持有者对实体的整体的观点倾向性 s。换句话说，基于2.1.4节对于观点五元组的定义，这一任务特别对GENERAL属性表达的情感进行分类：

$$(\_, GENERAL, s, \_, \_)$$

这里，实体、观点持有者以及观点的时间都已知，或与问题无关。

文档级情感分析可以按照 s 取值类型分为如下两种常见问题。如果 s 取类别值，例如褒义情感和贬义情感，它就是一个分类问题。如果 s 为数值或给定区间上的有序值，例如1~5星，那这个问题就变成了回归问题。

基于之前的讨论，我们会发现这个任务的假设其实具有很大的局限性，因为一篇文档中往往会包含多个观点，可以对多个实体进行评价，其中针对每个实体的观点倾向也可以不一样。观点持有者也许对一些实体持正面的态度而对另一些持负面的态度。在这种情形下，文档级情

感分类意义就不大了，因为对整个文档赋予一个情感基本没有什么用处。类似地，这一任务也对单个文档里多人观点的情况没有什么意义，因为他们的观点也可以不一样。例如句子，“Jane has used this camera for a few months. She said that she loved it. However，my experience has not been great with the camera. The pictures are all quite dark”。

假设3.1满足在线的产品和服务评论，因为在在线评论中，每则评论经常针对一个产品或服务，且也是由一个人写的。但是，这个假设对论坛讨论或者一篇博客就不一定成立了，因为一篇博客中作者可以对多个实体表达多个不同观点并加以比较。这也是为什么大多数研究者都使用在线评论来做情感分类或回归任务。

我们将在3.1节和3.2节中讨论情感分类问题，在3.3节中讨论预测情感评分的回归问题。文档级分类中，大多数现有的技术都是基于监督学习的，也有一些无监督学习的方法。对于情感回归任务来说，目前主要的方法是采用监督学习的技术。当然，还有一些拓展性的研究，目前主要集中在跨领域情感分类(或领域适应)和跨语言情感分类。我们将在3.4节和3.5节分别对其进行讨论。

尽管本章会详细介绍一些基本技术，但是由于提到了很多已经发表的文章，故主要还是一个综述的形式。现有大多数技术都是特征工程加机器学习算法在实际中的直接应用。但目前还
48 没有工作对于这些既有方法的有效性和准确性进行全面的、独立的评测和比较。

3.1 基于监督的情感分类

情感分类常被当作一个二类分类问题，将给定文本分为正面的和负面的情感。所用的训练和测试数据就是普通的产品评论。因为在线评论都包含评论者的评分，比如1~5星，所以依据这些评分，我们很容易得到正负例样本。4星或5星一般就是正面(褒义)的评论，而1到2星就是负面(贬义)的评论。大多数研究为了简便起见，并不使用中性分类(3星评论)。

情感分类本质上还是一个文本分类问题。但是，传统的文本分类主要是把文档分为不同主题，比如政治、科技或体育类。在这种分类中，主题词是重要的特征。然而在情感分类任务中，指示了正面或负面情感倾向的观点词或情感词更为重要，比如great、excellent、amazing、horrible、bad、worst等。本节我们会提到两类分类方法：(1)使用一个标准的有监督机器学习算法进行情感分类；(2)使用一个专为情感分类设计的分类方法。

3.1.1 基于机器学习算法的情感分类

因为情感分类是一个文本分类问题，所以任何监督学习方法都可以直接使用，比如朴素贝叶斯分类或支持向量机(SVM)(Joachims，1999；Shawe-Taylor and Cristianini，2000)。Pang等(2002)用这种方法分类影评。她们尝试了多种特征，但发现使用词袋(unigram)作为特征进行分类时，无论分类器选取朴素贝叶斯还是SVM，效果都非常好。

在后来的研究中，众多研究者尝试了非常多的特征和学习算法。和大多数监督学习应用一样，情感分类的关键还是抽取有效的特征。下面列出了一些特征样例：

词和词频。这些特征是带有词频信息的单独的词袋及与其相关的n-gram。这一特征在传统

的基于主题的文本分类中也经常使用。信息检索领域的 TFIDF 权重也可以应用在特征权重的计算上。和传统文本分类一样，这些特征在情感分类中也非常有效。

词性。每个词的词性(Part Of Speech，POS)是另一类特征。研究表明，形容词是观点和情感的主要承载词。因此一些研究者把形容词当作专门的特征进行特别处理。但是，我们可以把词性标签和 n-gram 作为特征混合起来使用。本书中我们使用宾州树库(Penn Treebank)词性标签集来表示不同的词性，如表 3-1 所示(Santorini，1990)。宾州树库的网址是 http://www.cis.upenn.edu ~ treebank/home.html。 49

表 3-1 宾州树库(Penn Treebank)词性标签

标签	描述	标签	描述
CC	Coordinating conjunction	PRP $	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb，comparative
EX	Existential *there*	RBS	Adverb，superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective，comparative	UH	Interjection
JJS	Adjective，superlative	VB	Verb，base form
LS	List item marker	VBD	Verb，past tense
MD	Modal	VBG	Verb，gerund or present participle
NN	Noun，singular or mass	VBN	Verb，past participle
NNS	Noun，plural	VBP	Verb，non-3rd person singular present
NNP	Proper noun，singular	VBZ	Verb，3rd person singular present
NNPS	Proper noun，plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP $	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

情感词和情感短语。情感词自然应该作为特征，毕竟它们就是那些在语言中表达了正面或负面情感的词语。例如，good、wonderful 和 amazing 是褒义词，而 bad、poor 和 terrible 是贬义词。大多情感词都是形容词或副词，但名词(如 rubbish、junk 和 crap)或动词(如 hate 和 love)也可以表达情感信息。除了单个词，也有一些情感短语或习语可作为特征，比如，cost someone an arm and a leg。

观点的规则。除了情感词和情感短语之外，还有很多文本结构或语言成分可以表示或隐含情感和观点，我们将在 5.2 节列举并讨论一些它们的表达方式。

情感转置词。有的表达可以反转文本中的情感倾向，比如把正面的情感倾向改变为负面的情感倾向。否定词是最重要的情感转置词。比如句子“I don’t like this camera”中的“like”是一个正面词，但它的情感倾向是负面的。此外，还有一些其他类型的情感转置词。我们将在 5.2、5.3、5.4 节中进行讨论。情感转置词需要小心处理，因为并不是有这类词，句子的情感

倾向就一定发生变化。比如“not only... but also”中的“not”就并没改变情感倾向。

句法依存关系。现有研究者也曾尝试从句法分析或句法依存树中取得词的依存关系特征。

50

在这个方向上已发表了大量文献，我们这里只能简要介绍一部分。

Gamon(2004)针对顾客的反馈数据进行分类。这种数据相比评论信息来说，常常会更短且含有更多的噪声。他们研究发现深层语言特征同 n-gram 类似，有助于提升分类效果。特征选择也很重要。通过使用微软研究院的 NLP 工具 NLPWin，他们获得了短语结构树并抽取了深层语言特征。这些特征包括 POS trigram，特定文本成分的长度信息(句子、从句、形容词性和副词性短语、名字短语等的长度)，句法树中每个成分基于上下文无关短语结构模式表示的成分结构(如 DECL:: NP VERB NP 表示一个由名词短语、动词、另一个名词短语依次组成的声明句)，带语义关系的词性信息(如“Verb-Subject-Noun”表示一个动词谓语接名词主语)，NLP-Win 工具提供的逻辑形式的特征，如谓语的及物属性以及时态信息。

Mullen 和 Collier(2004)介绍了一些可以和 n-gram 相结合的复杂特征。这些新特征分为三类：(1)利用词和短语的互信息(Pointwise Mutual Information，PMI)(Turney，2002)计算出情感值特征；(2)Osgood 等(1957)提出的有关形容词的三个因子值；(3)提及所评论实体的句子，在它附近或其中的属于 1、2 类的词或短语的情感值。Osgood 提到的三个因子是*强度*(强还是弱)、*主动性*(主动还是被动)和*评价*(好还是坏)，取值可从 WordNet 中所定义的语义关系中导出(Kamps et al.，2002)。这些新增特征比词根化后的词袋特征更加有效。我们将在 3.2.1 节中讨论 PMI 信息的详细情况。

Joshi 和 Penstein-Rosé(2009)在词袋特征之外，把依存句法关系和相关衍生特征应用到分类过程中。句子的依存句法分析结果是一组三元组$\{rel_i, w_j, w_k\}$，每个都包括了两个句子中的词及这两个词的句法关系，其中，rel_i是词w_j和w_k之间的依存句法关系。w_j通常指*首词*(head word)，w_k通常指*修饰词*(modifier word)。由这样的依存句法关系可以得到如 RELATION_HEAD_MODIFIER 形式的特征，这种形式的特征可以当作标准的词袋式二元特征，或基于计算频率的方式进行使用。比如“This is a great car”，在 great 和 car 之间有一个形容词限定(amod)句法关系，因此得到特征 amod_car_great。但是，这个特征针对性太强，只能用在汽车领域。我们可以对这一特征进行泛化，只使用首词的词性标签得到 amod_NN_great，这样就得到一个更通用的特征，可用到任意名词上。Xia 和 Zong(2010)把这个方法更进一步推广，使用 N 来代替 NN、NNS、NNP、NNPS 和 PRP，用 J 来代替 JJ、JJS 和 JJR，用 R 来代替 RB、RBS 和 RBR，用 V 来代替 VB、VBZ、VBD、VBN、VBG 和 VBP，用 O 来代替其他词性标签。他们丢弃了rel_i，因此 amod_car_great 就变成了两个特征 N_great 和 car_J。他们对传统的 bigram 词也采用同样的泛化策略。另外，对于分类模型，他们提出了一种集成模型(ensemble model)，提升了分类效果。Ng 等(2006)在较早的工作中也使用了依存关系(形容词－名词，主语－谓语，动词－宾语)作为特征，但他们并未对特征进行泛化。他们使用的特征有 unigram、bigram、trigram、情感词和评价对象词等。他们也发现，选择特征时，计算每个特征加权对数似然率作为特征权重，对于分类有较好效果。

51

Mejova 和 Srinivasan(2011)比较了不同类型的特征和特征选择策略。他们先测试了词干化(stemming)、词频和二元加权、基于否定的特征、n-gram 或短语的效果，然后尝试了基于频率的词汇剪枝、词性、词典的特征选择方法的效果。在不同大小的三个产品和影评数据集上，实验表明一些技术在大数据集上比在小数据集上更有用。对于大数据集，用互信息对特征进行排序，仅仅选用那些互信息值较高的少量特征训练出的分类器要比使用所有特征训练分类器的效果好。但是对小数据集就不是这样了。早期工作中，Cui 等(2006)对几个分类算法和最高到 6 阶的高阶 n-gram 做了一个评价。他们用了一个基于卡方检验的特征选择算法，并说明高阶的 n-gram 能获得更高的分类准确率。Bespalov 等(2011)的工作中也使用了高阶 n-gram，并使用了一个深度神经网络的方法来构建一个统一的判别式分类框架。Abbasi 等(2008)提出了一个基于遗传算法的特征选择算法，用于不同语言的情感分析。除了普通的 n-gram 特征之外，他们还用了文风特征，比如词汇丰富度和是否用到了功能词等。

对于微博情感分类，Kouloumpis 等(2011)使用了 4 种特征：(1)n-gram；(2)多角度问答(MPQA)主观性词典(Wilson et al.，2009)；(3)动词、副词、形容词、名词和其他词性的数量统计；(4)正面、负面、中性的表情符号以及缩写和强调(如全部大写或字母重复)的二元特征。

Pang 和 Lee(2004)没有使用整篇评论作为特征进行情感分类，他们提出只利用每条评论的主观部分作为特征进行情感分类。这些部分更可能包括观点和情感。为了找出评论中的主观成分，一个简单的办法是用标准的分类算法，把评论中的每个句子分为主观或客观，并独立对待每个句子。但是，文档中临近的句子具有一定的语义关系。考虑句子间的相近关系可以让算法充分使用文本连贯性：相邻的文本片段可能表达同样的主观性状态(主观或客观)，其他方面在临近句子间也具有一致性。为了考虑这种句子间相近关系，他们用图来表示一篇评论中的句子。设评论中的句子序列为 x_1，…，x_n。每个句子属于以下两类中的其中一类：C1(主观)和 C2(客观)。算法还用到下面两种信息： 52

- 个体的主观性得分 $ind_j(x_i)$。基于每个 x_i 内部的特征，判断其他属于 C_j 的非负估计。
- 联合得分 $\text{assoc}(x_i, x_k)$。对于 x_i 和 x_k 属于同一类的可能性的非负估计。

使用这些信息，可以构造了一个无向图 G，包括顶点 $\{v_1, \cdots, v_n, s, t\}$，其中，$v_1, \cdots, v_n$ 表示所有句子，s 和 t 分别表示正类和负类。算法向图中添加了 n 条边 (s, v_i)，每条边权重为 $ind_1(x_i)$，以及另外 n 条边 (v_i, t)，每条边权重为 $ind_2(x_i)$。最后我们加入了 $\binom{n}{2}$ 条边 (v_i, v_k)，每条边权重为 $\text{assoc}(x_i, x_k)$。分类任务就变为寻找这张图上的一个最小割的优化问题。个体的主观性得分 $ind_j(x_i)$ 由句子级主客观分类器给出，例如，朴素贝叶斯可以给出当前句子属于每个类的概率，这个概率就被用作 $ind_1(x_i)$ 和 $ind_2(x_i)$($=1-ind_1(x_i)$)。联合得分 $\text{assoc}(x_i, x_k)$ 则基于两个句子间的距离计算得到。最终的评论情感分类只使用最小分割算法所得到的主观句。这种分类方法比使用所有评论得出的分类结果更准确。

相关工作对情感的原因或证据(rationale)进行了标注，并利用这一信息来改善情感分类的效果。情感证据是指文档中由人工标注者或自动系统加上高亮的文本段，这段文本是支撑该文档中正面或负面情感的证据，也是表征主观信息的部分。Zaidan 等(2007)使用标注者来标注这

一信息。而 Yessenalina 等(2010a)使用基于情感词典的自动方法来标注情感的证据信息。Yessenalina 等(2010b)也试图找出带观点的或主观句子用于文档级情感分类。他们用一个基于结构化 SVM(Yu and Joachims，2009)的两级联合分类模型(句子级和文档级)，并直接优化文档级的分类。情感句子级的情感标签作为隐变量，因此在训练数据中不需要对其进行标注。McDonald 等(2007)也做了相似的工作，但他们的方法需要事先对训练集合中句子的情感信息进行标注。

Liu 等(2010)针对博客和评论情感分类任务，比较了不同的语言学特征。他们发现对博客数据的分类结果要比评论文本上的分类效果差很多，这是因为评论常常针对一个实体进行评价，而博客可以评价多个实体。其中对一些实体的观点可以是正面的，另一些又可能是负面的。随后，他们又研究了两种提升博客文本上情感分类准确性的办法。一种是基于信息检索的方法，找出每篇博客中与给定主题相关的句子，分类时不考虑那些与主题信息不相关的句子。
53 另一种方法是采用简单的领域适应技术。该方法先从评论领域训练几个分类器，然后利用训练得到的分类器对博客领域的文本进行分类，并把结果作为额外特征加入博客数据的训练过程之中。这种类型的扩展使得在博客数据上的分类准确性得以提升。

Becker 和 Aharonson(2010)从面向人类的心理语言学和心理物理学的实验角度，认为情感分类应当更加专注于文本的最后一部分(例如评论的最后一句)，但是这一观点并没有通过实验的结果得到证明。

现有方法都使用了 n-gram 特征(通常是词袋)，并使用信息检索领域中不同的方法来计算特征权重。Kim 等(2009)比较了不同的权重计算方法组合的效果，包括：PRESENCE(看是否出现二元指标)、TF(词频)、VS. TF(向量空间模型(VS)中的归一化词频)、BM25. TF(使用 BM25 中的归一化词频；Robertson and Zaragoza，2009)、IDF(逆文档频率)、VS. IDF(VS 中的归一化 IDF 值)、BM25. IDF(BM25 模型中归一化的 IDF 值)。结果显示 PRESENCE 类型的特征效果非常好。最好的组合是 BM25. TF 和 VS. IDF，但需要进行参数调整，并且其效果并没比 PRESENCE 好多少(大约提升了 1.5%)。

Martineau 和 Finin(2009)使用了一个新的特征权重计算策略，叫作 Delta TFIDF，实验表明该方法能够得到很好的结果。在这一方法中，一个词 t 和一篇文档 d 对应的特征值($V_{t,d}$)并不是该词在训练语料中 TFIDF 值，而是用下面的公式进行计算：

$$V_{t,d} = tf_{t,d} \times \log_2 \frac{N^+}{\mathrm{d}f_{t,+}} - tf_{t,d} \times \log_2 \frac{N^-}{\mathrm{d}f_{t,-}} = tf_{t,d} \times \log_2 \frac{N^+}{\mathrm{d}f_{t,+}} \frac{\mathrm{d}f_{t,-}}{N^-} \tag{3-1}$$

其中，$tf_{t,d}$(词频)是词 t 在文档 d 中出现的次数，$\mathrm{d}f_{t,+}$是训练集中包含 t 的正例文档数，N^+是训练语料中总的正例文档数，$\mathrm{d}f_{t,-}$是训练集中包含这个词的负例文档数，N^-是总的负例文档数。这个词频变换加强了在正负例中不均匀分布的词的重要性，而削弱了在正负例中均匀分布的词的重要性，更好地代表了文档中的词在情感分类时的重要程度。

Paltoglou 和 Thelwall(2010)做了一系列全面的实验，对不同的特征权重计算方法的有效性进行评估。所比较的方法包括 SMART 系统中的 TF 和 IDF 的变种计算方法(Salton，1971)、BM25 中相对应的变种的计算方法(Robertson and Zaragoza，2009)、SMART 系统中 Delta TFIDF

计算方法和 BM25 模型中 Delta TFIDF 计算方法。在每一个特种权重计算方法中，作者也使用了平滑技术。结果显示带平滑的 Delta 版本表现得比其他方法要好很多。

Li 等(2010)利用对个人(我、我们)和非个人(它、这件产品)的句子分别进行处理来提升情感分类效果。他们把个人句定义为句子的主语是(或者代表了)某人，而非个人句的主语则不代表人。他们分别使用个人句、非个人句和所有句子训练了 f_1、f_2 和 f_3 这三个分类器。这三个基本分类器分别使用他们的后验概率将其结果进行加权合并，用于最终类别的判别。

Li 等(2010)研究了否定词和其他情感转置词的处理方法，并将其用于改进文档级情感分类任务。他们用的办法与我们将在 3.2.2 节讨论的 Kennedy 和 Inkpen(2006)的做法不同，并不是基于词典的，而是基于监督学习的方法，这样不用明确识别单个具体的情感转置词。他们把文档中的句子分为有情感转置和未被转置两种。这里的分类并不需要人工标注数据。它仅仅利用了原始的文档级情感标签，以及一个特征选择方法。接着用这两种句子分别构造两个单独的情感分类器，然后再合并起来产生最终结果。Xia 等(2013)也提出了一个使用否定词进行情感分类的方法。

Qiu 等(2009)将基于词典和自学习的方法相结合(词典方法将在 3.2.2 节讨论)。简单来说，基于词典的方法使用了已给定情感词和情感短语来判别当前文档或句子的情感倾向。Qiu 等(2009)使用的算法包括两个步骤。第一步使用了基于词典的迭代方法，先初步用情感词典把一些评论进行分类，再用正负类样例的比例控制来迭代地判别其他评论的类别。第二步是在训练分类器的时候，利用第一步得到的分类结果作为训练数据。再用这个分类器去修正第一步中的分类结果。这个方法的优点在于它不需要标注数据，所以它实际上是一个使用监督技术的无监督方法，因此可以用到任何领域中。而基于训练语料的分类方法需要在所有领域中为训练所需的正负例进行人工标注。

Li 和 Zong(2008)尝试用多领域的训练数据来进行情感分类。他们用了两种融合方法，第一种是对多个领域内训练数据的特征进行融合，第二种则是对不同领域内训练得到的分类器进行融合。他们的结果表明分类器级别的合并要比单领域分类器(仅使用该领域的训练数据)的效果好。

Li 等(2009)提出了一个用于情感分析的非负矩阵分解模型。在这个模型中，$m \times n$ 的词－文档矩阵 X 被近似分解为三个因子，其中这三个因子规定了词和文档在第 k 个类中的近似归属关系，即 $X \approx FSG^{\mathrm{T}}$。$F$ 是一个 $m \times k$ 的非负矩阵，表示词在语义空间中的映射，即 F 的第 i 行代表第 i 个词属于第 k 个类的后验概率。G 是一个 $n \times k$ 的非负矩阵，表示文档在语义空间中的映射，即 G 的第 i 行表示文档 i 属于第 k 个类的后验概率。S 是一个 $k \times k$ 的非负矩阵，表示 X 的主成分分析后的状态。在二分类的情形下，$k=2$。我们可以通过分解得到 G 和 F 矩阵。G 告诉我们每篇文档的分类情况，F 告诉我们每个词的情感分类情况。因为不含任何初始知识，所以这其实是一个无监督模型。他们也做了一些监督实验，比如使用少量情感词和少量文档标签来训练分解模型。如果词 i 是一个褒义词，模型就设定 $(F0)_{i1}=1$；如果是贬义词，则设定 $(F0)_{i2}=1$。其中，$F0$ 是初始 F 矩阵。分解是一个基于三条更新规则的迭代过程。一些已知文档标签也可以采用类似的方法，如果第 i 个文档标记了正面情感的标签，则设定 $(G0)_{i1}=1$，否则设定 $(G0)_{i2}=1$。这种半

监督的分类方法也可产生很好的分类效果。

Bickerstaffe 和 Zukerman(2010)则研究了更一般性的问题，即对目标文档进行离散的、有序的情感多级打分，也就是对每个评论预测其评分星级的问题。因为类别是有序的，所以算法在分类时也考虑了情感类别间的相似度。

其他文档级情感分类工作包括：半监督学习和主动学习(Dasgupta and Ng，2009；Zhou et al.，2010；Li et al.，2011)，只标注特征而非文档(He，2010)，使用词向量捕捉词语的隐含语义以改善分类效果(Maas et al.，2011)，新闻陈述的褒贬分类(Scholz and Conrad，2013)，以及先进行词聚类降低特征稀疏度，再用聚好的词类作为特征构建分类模型(Popat et al.，2013)。Li 等(2012)使用主动学习，面对非平衡的训练数据进行情感分类。Tokuhisa 等(2008)研究了文本对话数据中的情绪分类问题。他们先进行三类(正面、负面、中性)情感分类，然后把得到的正面和负面言论分为十种情感类别。Aly 和 Atiya(2013)爬取了大量的阿拉伯文书籍评论(63 257 条)，并对其做了一些基本的情感分类和打分预测工作。

3.1.2 使用自定义打分函数的情感分类

除了使用标准的机器学习方法之外，研究者还提出了专门用于评论的情感分析技术。Dave 等(2003)提出的打分函数就是其中之一。它基于正面和负面评论词，主要包含如下两步：

第一步。用下面的等式为训练集中的每个词(unigram 或 n-gram)进行打分：

56

$$\text{score}(t_i) = \frac{\Pr(t_i \mid C) - \Pr(t_i \mid C')}{\Pr(t_i \mid C) + \Pr(t_i \mid C')} \tag{3-2}$$

其中，t_i是一个词，C 是一个类别，C'是它的补集，即非 C，$\Pr(t_i \mid C)$是词 t_i属于类别 C 的条件概率，通过将出现了 t_i的 C 类别文档数除以 C 类的总评论数计算得到。一个词的得分就是这个词对某个倾向类别相关度的度量，取值范围为 -1 到 1。

第二步。将一个新文档 $d_i = t_1 \cdots t_n$所有词的情感倾向性得分加起来，根据得分求得这篇文档的分类：

$$\text{class}(d_i) = \begin{cases} C & \text{eval}(d_i) > 0 \\ C' & \text{其他} \end{cases} \tag{3-3}$$

这里，

$$\text{eval}(d_i) = \sum_j \text{score}(t_j) \tag{3-4}$$

实验使用了包括 7 种产品超过 13 000 多条评论的数据集。结果显示选用 bigram(两个连续词)和 trigram(三个连续词)特征能够达到最高的准确率(84.6% ~88.3%)。实验中也没有使用任何词干归一化或移除停用词的操作。

他们也试验了一些其他分类方法，比如朴素贝叶斯、SVM、其他不同的打分函数以及词替换策略以提升泛化能力，比如：

- 产品名称用_productname 符号替换。
- 罕见词用_unique 符号替换。
- 类别专有的词用_producttypeword 符号替换。
- 数字符号用“NUMBER”替换。

他们也尝试了一些使用 WordNet、词干归一化、否定词、搭配词等语言学的处理方法。但是实验结果表明：这些操作并不那么有效，反而会降低分类的准确性。

3.2 基于无监督的情感分类

因为情感词常常主导了情感分类的结果，因此我们不难想到可以将它们以无监督的方式进行情感分类。这里主要讨论两种方法。一种是基于 Turney(2002)的做法，用那些可能表示观点的固定句法模板来分类。另一种是基于包含了褒义词和贬义词的情感词典的方法。

3.2.1 使用句法模板和网页检索的情感分类

Turney(2002)将每个句法模板看作一个带约束的词性标签序列(见表 3-2)。算法包括以下三步：

57

表 3-2 用以抽取两个词的短语的基于词性标签的模板

	第一个词	第二个词	第三个词
1	JJ	NN 或 NNS	任意
2	RB、RBR 或 RBS	JJ	非 NN 或 NNS
3	JJ	JJ	非 NN 或 NNS
4	NN 或 NNS	JJ	非 NN 或 NNS
5	RB、RBR 或 RBS	VB、VBD、VBN 或 VBG	任意

第一步。按照表 3-2 中所给的基于词性标记的模板在评论文本中抽取符合模板的两个连续的词。比如，模式 2 表示要抽取的两个连续词中第一个是副词，第二个是形容词，且后面跟着的词(不抽取)不能是名词。例如“This piano produces beautiful sounds”，beautiful sounds 就是要抽取的词，因为它满足模式 1。之所以用这些模式是因为具有 JJ、RB、RBR、RBS 词性标签的词经常用于表达观点和情感。名词或副词作为上下文是因为在不同的语境中，JJ、RB、RBR、RBS 等词表达的情感也可能不同。比如形容词(JJ)unpredictable 在一个汽车评论中可能暗含负面情感，例如“unpredicatable steering”，但在影评中又可能表示正面的情感，比如“unpredicatable plot”。

第二步。用互信息(Pointwise Mutual Information，PMI)来估计所抽取短语的情感倾向性(SO)：

$$\mathrm{PMI}(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right) \tag{3-5}$$

PMI 衡量的是两个词之间统计上的依存程度。此处 $\Pr(term_1 \wedge term_2)$ 是词 $term_1$ 和 $term_2$ 的真实共现概率，如果这两个词之间相互独立，则 $\Pr(term_1)\Pr(term_2)$ 是两个词的共现概率。短语

的情感倾向 SO 用它们与正面情感词 excellent 和负面情感词 poor 间的关联度计算得到：

$$\mathrm{SO}(\mathrm{phrase}) = \mathrm{PMI}(\mathrm{phrase},\text{“excellent”}) - \mathrm{PMI}(\mathrm{phrase},\text{“poor”}) \tag{3-6}$$

概率值可以通过将这两个词作为 query 提交给搜索引擎，并统计返回文档数的方法进行计算。对于每个查询，搜索引擎会给出相关文档的数目，我们称之为命中数。因此，通过合起来搜索两个词，以及分别搜索每个词，就能统计式(3-5)中的概率。Turney(2002)在其方法中使用了 AltaVista 搜索引擎，因为它有一个 NEAR 操作符可以限制两个词在文档中间隔的距离不超过 10 个词。将命中数记为 hits(query)，则式(3-6)可以写为：

58

$$\mathrm{SO}(\mathrm{phrase}) = \log_2\left(\frac{\mathrm{hits}(\mathrm{phrase}\ NEAR\text{“excellent”})\,\mathrm{hits}(\text{“poor”})}{\mathrm{hits}(\mathrm{phrase}\ NEAR\text{“poor”})\,\mathrm{hits}(\text{“excellent”})}\right) \tag{3-7}$$

第三步。给定一个评论，计算所有短语的 SO 值，如果平均 SO 值为正，则该评论的情感为褒义，反之为贬义情感。

最终在不同领域内分类的准确率，最高值出现在汽车评论领域内，为 84%，最低值出现在电影评论领域内，为 66%。

Feng 等(2013)使用不同语料比较了 PMI 和其他三种相关度计算方法。这三种指标分别为 Jaccard、Dice 和归一化的谷歌距离。所用语料为谷歌的索引页面、谷歌 Web 1T 5-gram 数据集、维基百科和推特。他们的实验结果表明在推特语料上计算 PMI 的最终效果最好。

3.2.2 使用情感词典的情感分类

另一种无监督的方法是基于词典的情感分类方法。其主要特点是，分类是基于一个包含已标注的情感词和短语的词典。这个词典称为*情感词典*或*观点词典*，其中包括了情感词和情感短语的情感倾向性和情感强度。每篇文档的情感得分还需要结合情感加强词和否定词来进行计算(Kennedy and Inkpen，2006；Taboada et al.，2006，2011)。这种方法之前用于基于属性的情感分类(Hu and Liu，2006)和句子级的情感分类(Kim and Hovy，2004)。这种方法对于表达了正面情感的文本表达(词或短语)都赋予了一个正的 SO 值，而每个表达了负面情感的文本表达都赋予一个负的 SO 值。

这种方法的基本形式是将文档中所有情感表达的 SO 值求和。若和的值为正，则该文档就被判定为包含正面的情感；若和的值为负，则该文档就被判定为包含负面的情感；若和的值为 0，该文档就被判定为包含中性的情感。该方法有很多变种，主要的不同在于每个情感的表达被赋予什么样的值，否定词如何处理，是否考虑新增信息等。Hu 和 Liu(2004)与 Kim 和 Hovy(2004)的方法类似，为每个正面情感表达赋予 +1，为每个负面表达赋值 -1。如果出现否定词，如 not 和 never 等，则将 SO 值取反。例如，good 的情感 SO 值是 +1，而 not good 是 -1。Polanyi 和 Zaenen(2004)证明了除否定词之外的其他因素也会影响特定情感表达倾向性的正负性。这些因素又叫*情感转置词*(sentiment shifter)(或价转移(valence shifter)；Polanyi and Zaenen，2004)。情感转置词是可以改变另一个表达的 SO 值的词。实际上，情感转置词要比 Polanyi 和 Zaenen(2004)中列举的多得多。我们将在 5.2 ~ 5.4 节中详细讨论这一问题，同时介绍除了情感表达之外传递或暗示情感的方式与方法。

Kennedy 和 Inkpen(2006)实现了 Polanyi 和 Zaenen(2004)的部分想法。除了改变或反转情感 59
的否定词外，他们也研究了前面提到的情感加强词和减弱词。这些词可以改变所表达情感的强度。情感加强词会增加正面或负面情感的强度，而减弱词则削弱之。例如句子“This movie is very good”，短语 very good 就比 good 具有更强的情感，而句子“This movie is barely any good”中，barely 就是一个减弱词，使得句子情感非常负面。为了考虑加强词和减弱词，该论文将所有的正面情感赋值为 2。如果前面在同一从句中出现加强词，就赋值为 3。如果前面有同一从句的减弱词，值就为 1。同样，负面情感词的默认情感打分为 -2，如果前面有减弱词或加强词就分别赋值 -1 或 -3。

Taboada 等(2011)对于上述方法进行扩展，进一步考虑更精细的情形。每个情感表达的 SO 值范围是 -5(极度否定)到 +5(极度肯定)，0 值除外。每个加强词和减弱词都有一个或正或负的权重百分比。例如，slightly 是 -50，somewhat 是 -30，pretty 是 -10，really 是 +15，very 是 +25，extraordinarily 是 +50，而(the)most 是 +100。如果 excellent 的 SO 值为 5，那么 most excellent 的 SO 值就即为 5 ×(100% +100%) =10。加强词和减弱词从最靠近情感表示开始，顺序地逐步加入 SO 值的计算：如果 good 的 SO 值为 3，则 really very good 的 SO 值就为 [3 ×(100% +25%)] ×(100% +15%) =4.3。情感加强和减弱的形式主要有两种：一种是有 SO 值的形容词带副词修饰(如 very good)；另一种是有 SO 值的名词带形容词修饰(如 total failure)。除了副词和形容词之外，Taboada 等(2011)还用到了其他词性的加强词和减弱词：数量词(a great deal of)、全部字母大写、感叹号标记，以及语篇连接词 but(给出更多显著信息)。

在很多情况下，当遇到否定词时，简单反转 SO 值会有问题。比如 excellent 是一个 SO 值为 +5 的形容词；如果将其与否定词搭配就变成 not excellent，其 SO 值为 -5，这与 atrocious 大相径庭。实际上，not excellent 看起来比 not good 还要偏向正面的情感倾向，给一个 -3 之类的值就可以。为了捕获这样的实际感觉，SO 值会向着相反极性方向移动一个固定值(如 4)。因此一个情感倾向值为 +2 的形容词，当遇到一个否定词时，其情感倾向会被否定到 -2。但一个 SO 值为 -3 的形容词(如 sleazy)，当遇到否定词时，其 SO 值只会变得稍微正面一点(即 +1)。下面给出一些例子：

a. 它并不极好(5 -4 =1)，但也并不太差(-5 +4 = -1)。
b. 我必须承认这并不差(-3 +4 =1)。
c. 这张 CD 不那么难听(-5 +4 = -1)。

这背后的基本思想是：在某种程度上，如果我们没有暗示说得到一个弱正面情感的表达，的确很难确定否定词对一个强正面词的情感倾向进行了反转。因此，否定词从某种意义上说也变成了一个情感减弱词。 60

Kennedy 和 Inkpen(2006)还注意到，基于词典的情感分类器通常会偏向正的情感倾向。为了抹平这种误差，Taboada 等(2011)为所有负面的最终 SO 值(计算过其他修饰词之后)增加了 50%，从而为相对少见的负面词赋予了更多的权重。

另外还有很多标记词表明句子中一些词不适合做情感分析。这类词一般暗示上下文是不真

实的，又被称为假定的、虚拟的语气。其标记词包括：情态动词、条件标记词(if)、负面词(any 和 anything)、确定性(主要是强度)动词(expect 和 doubt)、疑问句、引号中的词(可能是主观性的词，但不一定会反映作者的观点)。分析过程中，在这些虚拟词作用范围外(如同一从句)的词的 SO 值将不会被考虑。这种策略称为虚拟阻断。这并不是说这些句子或从句不带情感。实际上很多这类句子确实含有情感，例如“Anyone know how to repair this lousy car?”然而，要可靠地区分这类句子是否表达了情感十分困难，因此只能忽略。我们将在讨论句子级和属性级的情感分类时进一步讨论这些问题。

除了上述提到的方法之外，针对一些特定应用也有一些人工的方法。比如 Tong(2001)介绍了一个生成情感时间轴的系统。这个系统追踪在线的电影讨论，展示了一个正面和负面评论(Y 轴)随时间(X 轴)变化的图。评论的消息根据特定短语进行匹配，从而对这些消息进行分类，之后我们可以获得作者对某一部电影的观点倾向性，比如 great acting、wonderful visuals、uneven editing、highly recommend it 等。这些短语通常是人工编纂的。因此所得词典只是针对这个领域，当面对一个新领域时，又需要人工编纂一个新词典。

我们发现，如果一个领域有大量标注数据，那么监督学习通常会有更好的分类准确性，因为它自动学到了特定领域的情感表达。基于词典的情感分类方法识别特定领域表达就没那么容易，除非有一个可以发现这些表达并自动确定其倾向性的算法。这方面已经有一些工作(Zhang and Liu，2011a，2011b)，但还远非成熟。监督学习也有它自身的弱点。最主要的就是用一个领域数据训练得到的分类器不能用到另一个领域(见 3.4 节)。因此，每个应用领域都需要训练数据才能获得精准的分类。这时，基于词典的方法不需要训练数据，因而相对于监督学习的方法更具优势。

3.3 情感评分预测

对有的应用来说，只把观点的倾向性二分为正面或负面是不够的，用户还需要分析出正面
61 情感和负面情感的强度。为此，研究者尝试了一些预测评论打分(如 1～5 星)的研究(Pang and Lee，2005)。由于情感的打分是一个序数，因此这个问题一般被形式化为回归问题。但并不是所有研究者都采用回归的方法解决这一问题。Pang 和 Lee(2005)尝试了 SVM 回归、基于一对多(One-Versus-All，OVA)策略的 SVM 多类分类，以及被称作度量标注的元学习的不同方法。结果显示基于 OVA 的分类比其他两者效果差得多。这一结论也比较容易理解，毕竟数值化的打分不是类别值。Goldberg 和 Zhu(2006)对这个方法进行了改进，将评分预测问题建模为基于图的半监督学习问题，既带有标注数据也有未标注数据。未标注数据是需要进行预测打分的评论。在这个图中，每个节点是一个评论文档，两个节点之间连接的权重是两个文档的相似度。相似度很高表明两个文档的评分很相近。作者尝试过不同的相似度度量方法。他们还假设在算法初始化时，已有一个单独的学习者对于未标注的文档已经预测了其评分。这个图方法通过求解优化问题，基于初始的评分以及图上边的权重，对于文档的评分进行修正，从而使得整个图上的评分更平滑。

Qiu 等(2010)修改了传统的词袋表示，引入了文档的“观点袋”(bag-of-opinion)表示，以捕获观点中 n-gram 的情感强度。他们认为每个观点都是一个三元组，包含情感词、修饰词和否定词。例如，“not very good”中的 good 是情感词，very 是修饰词，not 是否定词。对于两类

情感分类，观点修饰词不太重要，但对于评分预测就很有用了，否定词也是一样。他们利用有约束的岭回归(ridge regression)算法在领域无关语料中学习观点强度或情感得分。其中关键点是对于已有观点词典和评论评分的运用。为了把训练好的回归模型用到一个新的领域的应用当中，基于观点评分，他们设计了一些统计量，并将其作为额外特征与标准的词袋一起进行新领域内评论的评分预测。在这之前，Liu 和 Seneff(2009)在通过句法分析所得到的句子的层次化语义结构表示的基础上，提出了一个抽取“副词 - 形容词 - 名词”短语的方法(例如，“very nice car”)。他们没有用到机器学习的方法，而是用了一个启发式的算法，基于已计算的每个词、短语的情感得分，通过计算形容词、副词、否定词等对于整句的情感打分的贡献，获得整篇评论的情感打分。

Snyder 和 Barzilay(2007)没有对整篇评论进行情感打分预测，而是研究了针对其中每个属性的评分预测问题。我们可以简单地用回归和分类的算法来做，但这个方法没有用到针对不同属性的用户观点之间的依赖关系，而这一信息往往对提升预测准确率非常有用。针对这一问 62
题，这篇文章提出了两个模型：基于属性的模型(作用于每个属性)和基于一致性的模型(用于建模属性间的评分的一致性)。在学习过程中同时考虑两个模型，其中训练所用特征是每个评论的 unigram 和 bigram 之类的词汇特征。

Long 等(2010)用了与 Pang 和 Lee(2005)类似的方法，但他们使用了贝叶斯网络分类器来预测每个属性的评分。为了得到较好的准确率，他们没有对每一篇评论都进行预测，而是只关注那些对于属性进行了全面评价的评论，并从中预测出针对属性的评分，这是由于其他评论没有足够的信息。评论的选择是基于 Kolmogorov 复杂度的信息度量方法。然后利用机器学习算法对所选评论的属性进行评分预测，训练用到的特征仅来自于属性相关的句子。属性的抽取采用与 Hu 和 Liu(2004)类似的做法。

3.4 跨领域情感分类

情感分类对训练数据所属领域异常敏感。一个领域的观点文档所训练得出的分类器用到别的领域效果通常都很差，这是因为不同领域的词甚至语言构成可能都大不一样。更糟的是，一个词在不同领域可能连情感倾向性都不一样。因此需要领域适应(domain adaptation)或者迁移学习(transfer learning)的技术。已有研究主要基于两种前提条件：一种是需要新领域内也有少量标注数据(Aue and Gamon，2005)；另一种则不需要新领域的标注数据(Blitzer et al.，2007；Tan et al.，2007)。有标注的原始领域一般称为源领域，待测试的新领域叫作目标领域。

Aue 和 Gamon(2005)在缺少目标领域大量标注数据的条件下，给出了一种情感迁移分类算法。他们尝试了 4 种策略：(1)混合其他领域的标注数据用于训练，然后在目标领域测试；(2)像(1)一样训练分类器，但是仅使用目标领域中存在的特征；(3)使用多领域分类器的集成(ensemble)，并在目标领域测试；(4)结合使用目标领域的少量标注数据和大量未标注数据(这是传统的半监督学习方式)。前三种策略使用 SVM 作为分类器，第四种策略使用期望最大化(Expectation Maximization，EM)的半监督学习(Nigam et al.，2000)。他们的实验表明，第四种策略的效果最好，这是由于这一策略使用了目标领域的标注和未标注数据。

Yang 等(2006)提出了一种基于特征选择的简单策略，用于句子级分类的迁移学习任务。 63

他们先用了两个领域内充分的标注数据来选出两个领域都有效的特征。这些特征被视为领域无关的，用它们构造的分类器被认为能用到任何目标领域。Tan 等(2007)提出了另一个简单的方法，先用源领域数据训练一个基本分类器，再用它标记一些目标领域有代表性的样本。基于这些目标领域内选出的样本，再训练一个分类器用于目标领域的分类。

Blitzer 等(2007)早期提出结构化对应学习的方法(Structural Correspondence Learning，SCL)(Blitzer et al.，2006)来做领域自适应。给定源领域的标注数据和未标注数据，以及目标领域的未标注数据，SCL 算法先选择了两个领域都频繁出现的 m 个特征，它们对源领域预测效果最好(在文章中就是源标签互信息具有最大值的那些特征)。这些中轴特征代表两个领域共享的特征空间。其次，SCL 计算每个中轴特征和每个领域的非中轴特征的相关性，得到相关矩阵 $\boldsymbol{W}$。它的第 i 行是表示第 i 个中轴特征对每个非中轴特征的相关度向量。直观上说，相关数为正就表示在该领域的非中轴特征与中轴特征正相关。这就建立起两个领域的特征对应关系。之后使用 SVD 分解计算 $\boldsymbol{W}$ 的一个低维线性近似 θ(前 k 个特征向量的转置)。最终用来训练和测试的特征就是原始特征 $\boldsymbol{x}$ 及其变换后的 $\theta\boldsymbol{x}$，即 k 个实值特征。基于这一特征集和源领域的标记数据所构建的分类器可以在源领域和目标领域都取得很好的结果。

Pan 等(2010)在更高层次提出了一个与 SCL 想法相似的方法。这个算法的前提是只有源领域有标记数据，而未标记领域只有未标记数据。它通过谱特征对齐(Spectral Feature Alighment，SFA)算法将不同领域的领域有关词聚为多个簇，簇之间通过领域无关词连接。领域无关词与 Blitzer 等(2007)所用的中轴词相似，因此可以用相似的办法选取。SFA 先构造了领域无关词和领域有关词之间的二分图。领域有关词若与领域无关词在文档或窗口中共现，则它们之间就有一条边，边上的权重是它们的共现频率。接下来用谱聚类算法在二分图上将领域有关词和领域无关词聚为特征簇。这里的想法是，如果两个领域相关词连接了很多相同的领域无关词，则它们就可能聚为一类。同样，如果两个领域无关词与很多相同的领域相关词相连，则它们也应该聚为一类。在最终的跨领域训练和测试中，这些簇和原始特征集相结合组成新的特征集，所有的样本都用新的特征集来表示。

64

基于同样的想法，He 等(2011)使用联合主题建模来识别两个领域的观点主题(与前面说的簇相似)，并建立不同主题的链接。所得话题涵盖了两个领域，被直接当作新增特征来扩充原始的特征集合。Gao 和 Li(2011)也用了主题建模，试图基于两个领域的词对应关系和词共现信息来找到两个不同领域之间一个共同的语义空间。基于这个语义空间就可以训练出一个目标领域的分类器。Bollegala 等(2011)提出一种方法，该方法可以用多个源领域的标注和非标注数据，构造一个基于情感信息的同义词库，这样便于找到不同领域中具有相似情感的词之间的联系。这个同义词库可以用来扩展原始特征向量，用于训练一个两类情感的分类器。Yoshida 等(2011)通过识别领域相关、领域无关的情感词，给出了一种从多个源领域到多个目标领域的情感迁移方法。Andreevskaia 和 Bergler(2008)使用了两个分类器的集成(ensemble)进行领域迁移。第一个分类器使用词典构造，第二个使用少量领域内数据训练得到。

Wu 等(2009)提出一种基于图的方法，主要是基于相似度图上标签传播的思想(Zhu and Ghahramani，2002)来做领域迁移。这个图中每个文档是一个节点，每条边表示文档间的相似度，边上的权重可以用余弦相似度进行计算。每个源领域的文档标签值初始化为 +1(正面)或

-1(负面)，每个目标领域的标签值则可以基于一个普通情感分类器训练得到。之后算法就进行迭代，针对每个目标领域 i，通过找出其 k 个源领域的最近邻和 k 个新领域的最近邻，将这些邻居的标签得分进行线性组合，以此来更新当前样本 i 的标签值。整个算法当标签值收敛时停止迭代。目标领域的情感倾向标签依据它们的文档标签得分进行确定。Ponomareva 和 Thelwall(2012)比较了基于图的方法和其他几个现有最好方法。结论显示，基于图的方法在领域适应问题上给出了很有竞争力的结果。

Xia 和 Zong(2011)发现，在不同领域，一些词性标签特征常常是领域相关的，而另有一些是领域无关的。基于这个观察，他们提出基于词性的组合模型来结合不同的词性特征，以改善分类性能。

3.5 跨语言情感分类

跨语言情感分类就是对多种语言的观点文档进行情感分类。这一任务主要有两个动机。第 65
一，不同国家的研究者希望能建立针对本国语言的情感分析系统。但是，已有大部分研究工作都是针对英语的，其他语言没有太多资源或工具可以迅速建立一个不错的情感分类器。于是就有了一个很自然的想法，是否可以利用机器翻译和现有的英语情感分析系统资源和工具，来构建其他语言的类似系统。第二个动机是很多公司都希望了解和比较不同国家的消费者对他们产品的观点。如果他们有一个英语的情感分析系统，便希望通过翻译手段，尽快建立一个其他语言的情感分析系统。

很多研究者都研究过这个问题。很多现有工作都主要关注于文档级别的情感分类，以及句子级别的主观性和情感分类。除了 Guo 等(2010)的工作之外，目前针对属性级别的研究工作还不多见。本节我们主要关心文档级别的跨语言情感分类任务。4.6 节会研究句子级别的跨语言情感分类问题。

Wan(2008)利用英语的情感资源做中文的评论分类。第一步是用多个翻译引擎把每条中文评论都翻译为英文，得到多个不同的英语版本。之后他们用基于词典的情感分类方法对每份英语翻译进行分类。情感词典包括正面情感表达(词和短语)、负面情感表达、否定的文本表达、情感加强词。最后把评论中所有情感表达在考虑了情感加强和情感否定之后的情感得分加起来，若最终分值小于 0，则该评论的情感就是负面的，否则就是正面的。然后，他们用不同的组合方法对每个评论不同翻译的得分进行组合(平均、最大值、加权平均、投票等)，以获取该篇评论的情感分类结果。如果有中文情感词典，那也能对原始中文评论使用同样的基于词典的情感分类办法获取该评论的所属情感类比，且可以和英文翻译的结果进行结合。他们的结果显示组合技术非常有效。Brooke 等(2009)也尝试了用一个英语到西班牙语的翻译引擎获取该评论的西班牙语翻译结果，再用基于词典或机器学习的办法对目标语言进行文档级情感分类。

Wan(2009)提出了一种用英语标注语料进行协同训练的方法，通过监督学习的手段进行中文评论分类，而该方法没有用到任何中文资源。在训练时，输入为一组有标注的英文评论和一组未标注的中文评论。有标注的英文评论会被翻译成中文，而未标注的中文评论会被翻译成英文。于是每条评论就都有一个英文版本和中文版本。英文特征和中文特征被看作同一评论的两个独立而冗余的不同表示。之后该方法使用基于 SVM 的协同训练方法训练出两个分类器，之 66

后再合并为一个分类器。测试时每条未标注的中文评论都先翻译为英文，再用上述训练得到的分类器将其分为正面的情感或负面的情感。Wan(2013)使用了一个协同回归的方法预测跨语言评论的打分，该方法也用了协同训练的想法。

Wei 和 Pal(2010)用基于迁移学习的方法进行跨语言情感分类。由于机器翻译经常出错，为了降低翻译所引入的噪声，他们提出了一种方法，基于 SCL 方法(Blitzer et al.，2007)，选出由两种语言(中文和英文)所共有的核心特征。为了减少数据和特征的稀疏性，他们向搜索引擎提交查询，试图找到那些和核心特征相关的特征，然后用这些新发现的特征构建更多伪样例用以训练分类器。

Boyd-Graber 和 Resnik(2010)对于有监督的 LDA 主题模型(SLDA)(Blei and McAuliffe，2007)进行扩展，用于预测多语言评论的打分。SLDA 可以在主题建模时考虑用户打分。这个扩展模型 MLSLDA 同时从多语言文档中挖掘主题信息，所得到的主题在多个语言里具有一致性。为了把不同语言的主题词关联起来，这个模型还用到了不同语言的 WordNet 以及已对齐的词典信息。

Guo 等(2010)使用一个基于主题模型的方法把多语言属性的不同表达进行聚类，从而用于对不同国家中基于属性的情感分析结果的比较(见 9.1 节)。

Duh 等(2011)针对跨语言情感分类问题进行了研究，他们的分析认为，领域不匹配不是由于机器翻译的错误引起的。哪怕机器翻译结果完全正确也会造成跨语言分类准确率下降。他们还认为跨语言的自适应问题和其他(单语言的)自适应问题有着本质不同，因此应当考虑新的自适应算法。

3.6 文档的情绪分类

现在我们转而考虑情绪分类问题，这一任务相对于情感分类任务要难得多。这是因为：(1)情绪有更多的类别，即情感和心情的类型；(2)不同类型的情绪和心情有非常多的相似之处，难以区分。由于难以对情绪和心情进行区分(Alm，2008)，因此在本节我们不对它们进行区分。

现有的文档级情绪分类大多都是采用监督学习的方法。例如，Mishne 和 de Rijke(2006)针
67 对来自 LiveJournal. com 的博客数据进行情绪分类。在 LiveJournal. com 上，博主可以为他们的博文加上心情标签，因此这些博文就能用于监督学习。主要的特征是那些可代表每种情绪的词(或 n-gram)。这些词可以这样计算：每种情绪类型 m 生成两种概率分布 θ_m 和 $\theta_{\neg m}$。θ_m 是所有标记了情绪类型为 m 的博文中词的分布，而 $\theta_{\neg m}$ 是剩余博文中的词的分布。θ_m 中的所有词都用他们的对数似然值进行排序，并与 $\theta_{\neg m}$ 中该词的对数似然值进行比较，这样得到了针对情绪类型 m 特征词的有序列表。当对每种情绪都完成了这一操作，就可以选出每个有序列表的前 N 个词构成针对每种情绪的代表性特征集。其他类型的特征也有使用，比如博文发布于一天中的时间、发布日期、是否是在周末发布等。对于模型构建，该方法用了 Pace 回归算法(Wang and Witten，1999)。

Lin 等(2007)用雅虎提供的中文新闻文章进行情绪识别。在雅虎的新闻网页中，读者可以

基于自己感受到的情绪对其进行投票。基于这样的数据，算法使用 SVM 进行监督学习。总共用到了四种特征集。第一种是基于汉字字符的 bigram 特征。第二种包括中文分词后产生的所有词。第三种是文章的元数据，比如新闻记者、新闻分类、新闻事件地点、发布时间、新闻社名称等。第四个集合是词的情绪类别，词的情绪类别从作者之前构造的情绪词典中获得(Yang et al.，2007)。

Strapparava 和 Mihalcea(2008)也用到了监督学习方法。他们使用的是朴素贝叶斯分类。另一种监督学习方法是基于流形学习的(Kim et al.，2013)。这里的学习算法与前面那些文章的不同之处在于，它们的方法把情绪预测任务看作一种离散标签的多分类问题。这篇文章假设数据满足连续的情绪流形，故他们的方法在本质上是不同的学习范式。

其他的情绪研究工作主要都与情绪词典的构建有关，比如 WordNet-Affect(Strapparava and-Valitutti，2004)就是基于 WordNet 构建的，而 Mohammad 和 Turney(2010)所构建的词典则是通过众包的方式构建的。Mihalcea 和 Liu(2006)以及 Mohammad(2011)也用不同种类的在线文本做了一些情绪分析工作(不是情绪分类)。

3.7 小结

文档级的情感分类目标是检测整篇文档的整体观点和情感。已有很多研究者对其做了广泛的研究。但是，在这个级别上进行分类有如下两个缺陷：

- 它不考虑情感或观点所评价的对象。尽管对评论文本来说，这种方法已经足够，但那是因为一条评论通常只评价一个实体而已。对非评论，比如论坛讨论、博客、新闻等数据，这种方法就捉襟见肘了。这是因为这些帖子都会同时评价多个实体，并使用比较句比较每个提到的实体。因为论坛帖子不像评论，它可能会给出一些产品描述，而不管对该实体持有什么样的观点，所以很多时候难以确定一篇帖子是不是评价了用户感兴趣的实体，或者是不是表达了观点。文档级别的情感分类不能完成如此精细的任务，这需要深度的自然语言处理，而不仅仅是进行文本分类。实际上，因为几乎所有在线评论都有用户评的星级，因此这一类评论数据并不需要情感分类。在实际应用中，需要情感分类的往往是论坛讨论和博客，以便确定人们的观点。 68
- 即使知道一篇文档只评价了一个实体，大多数应用中用户还是想知道更多细节，比如消费者喜欢产品的哪方面、不喜欢哪方面。在一篇典型的包含观点的文档中，这些细节都是有的，但文档级别的情感分析不能为用户提取这些内容，这些细节也可能对用户做决策非常重要。比如一款获得了全好评的相机(4 星或 5 星)，但一些评论者提到它的电池续航时间短。如果一个潜在的消费者需要更长的续航时间，那么哪怕每篇相机评论都是正面的，他可能也不会买。 69