
情感分析
Sentiment Analysis
徐冰　副教授
哈尔滨工业大学

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

第三章 文档级情感分类

- **3.1** 基于有监督的情感分类
- **3.2** 基于无监督的情感分类
- **3.3** 情感评分预测
- **3.4** 跨领域情感分类
- **3.5** 跨语言情感分类
- **3.6** 文档的情绪分类

第三章 文档级情感分类

- 文档级情感分类（**Document sentiment classification**）
  - 任务目标：将一篇给定观点的文档根据所持观点为正面或负面进行分类。
  - 正面或负面的观点又称为 情感倾向性 或 极性
  - 将一篇文档看做整体，不研究文档中具体的实体或属性。
- 方法：传统的文本分类方法
- 定义：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的实体e，且只包含一个观点持有者h的观点。
  - 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。

文档级情感分类

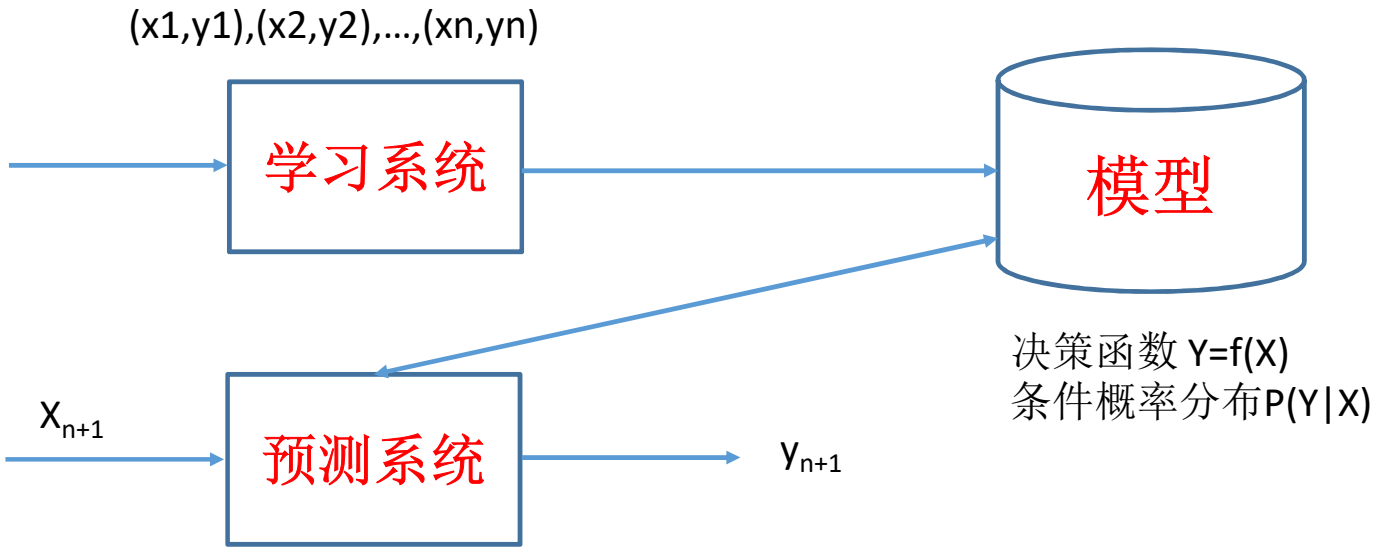
- 存在的问题：一篇文档中可能会包含多个观点，可以对多个实体进行评价，每个实体的观点倾向也可能不同。
  - *eg. Jane has used this camera for a few months. She said that she loved it. However, my experience has not been great with the camera. The picture are all quite dark.*
- 比如在线评论，每个评论针对一个产品或服务，可以做情感分类或回归任务。

基础知识介绍

- 统计学习(Statistical Learning):
  - 关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。统计学习也称为统计机器学习。
  - 统计学习的特点：
    - 以计算机及网络为平台，是建立在计算机及网络之上的
    - 以数据为研究对象，是数据驱动的学科
    - 目的是对数据进行预测和分析
    - 以方法为中心，构建模型并应用模型进行预测与分析
    - 统计学习是概率论、统计学、信息论等多领域的交叉学科。
  - 统计学习包括：监督学习，无监督学习，半监督学习及强化学习。

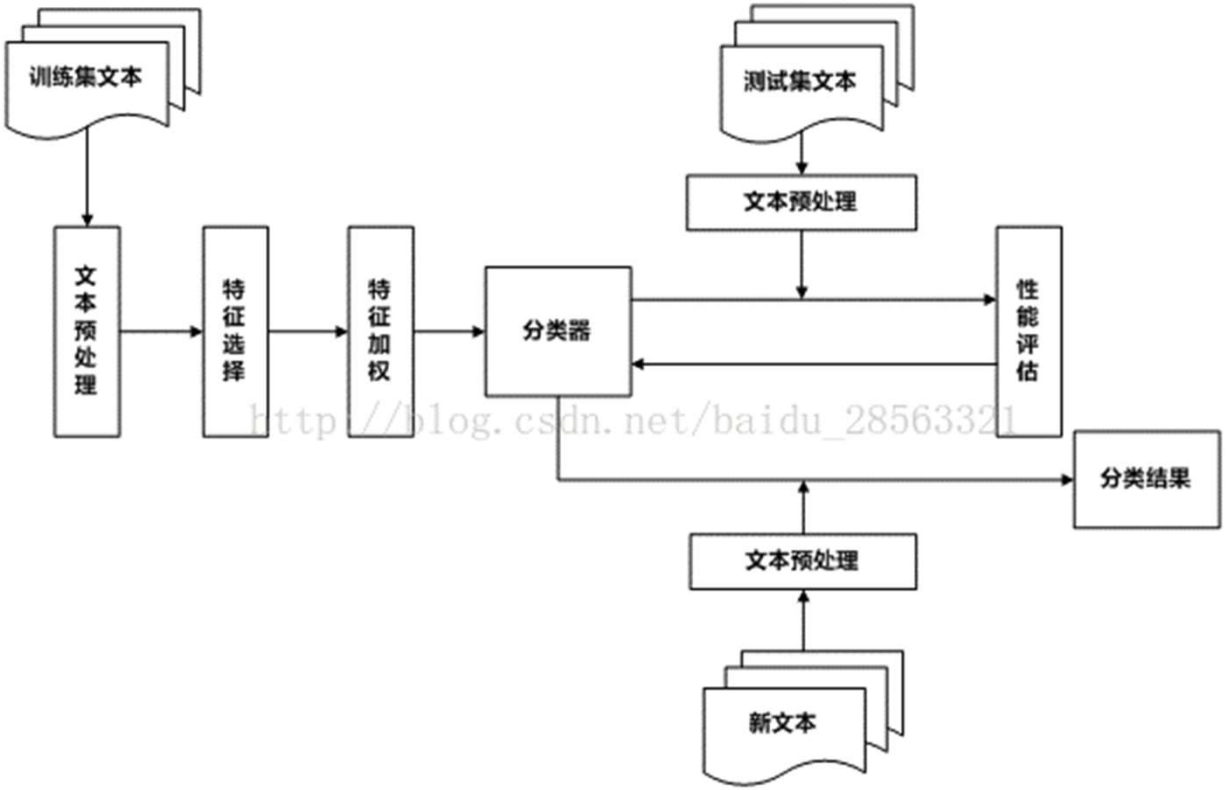
监督学习问题

- 监督学习：利用训练数据集学习一个模型，再用模型对测试样本集进行预测。由于在这个过程中，需要训练数据集，而训练数据集是人工给出的，所以称为监督学习。



- 评价分类器性能指标：
  - 分类准确率（**accuracy**）:对于给定的测试数据集，分类器正确分类的样本数与总体样本数之比
  - 二分类问题常用评价指标：精确率（**precision**）、召回率（**recall**）、**F**值
    - 通常以关注的类为正类、其他类为负类
    - 分类器在测试数据集上的预测或正确或不正确，四种情况
      - **TP**-将正类预测为正类数
      - **FN**-将正类预测为负类数
      - **FP**-将负类预测为正类数
      - **TN**-将负类预测为负类数
    - 精确率 $\mathrm{P}=\frac{TP}{TP+FP}$　　召回率 $\mathrm{R}=\frac{TP}{TP+\mathrm{FN}}$　　**F1**值 $\mathrm{F1}=\frac{2TP}{2TP+FP+\mathrm{FN}}=\frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}$

文本分类流程



3.1 基于有监督的情感分类

• 3.1.1 基于机器学习算法的情感分类

- 二类分类问题：给定文本分为正面和负面的情感
- 已有研究所用的方法：
  - 朴素贝叶斯分类器
  - 支持向量机（SVM）
  - 最大熵分类器
  - ……
- 采用的特征：
  - Unigram（一元文法）
  - Bigram（二元文法）
  - ……

unigram 形式为：哈/尔/滨/工/业/大/学
bigram形式为：哈尔/尔滨/滨工/工业/业大/大学
trigram形式为：哈尔滨/尔滨工/滨工业/工业大/业大学

• 3.1.1 基于机器学习算法的情感分类

- 情感分析的关键在于抽取有效的特征：

（1）词和词频：带有词频信息的**Unigram**和**n-gram.**

（2）词性：**POS**，形容词是观点和情感的关键词。

- ✓英文：宾州树库（**Penn Treebank**）
- ✓中文：人民日报语料标注

（3）情感词和情感短语

- ✓褒义词：**good, wonderful**
- ✓贬义词：**bad， poor**

（4）观点的规则：除了情感词以外，很多文本结构或语言成分可以表示或隐含观点

（5）情感转置词：否定词

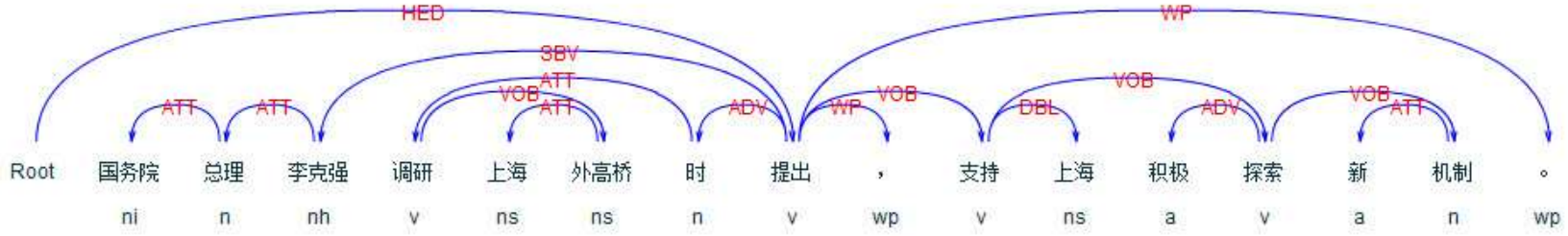
（6）句法依存关系：研究工作较多

语言基础分析

- 中文分词
  - 中文分词 (Word Segmentation, WS) 指的是将汉字序列切分成词序列。 因为在汉语中，词是承载语义的最基本的单元。分词是信息检索、文本分类、情感分析等多项中文自然语言处理任务的基础。
    - 例如：*国务院总理李克强调研上海外高桥时提出，支持上海积极探索新机制。*
    - 分词结果：国务院/ 总理/ 李克强/ 调研/ 上海/ 外高桥/ 时/ 提出/ ，/ 支持/ 上海/ 积极/ 探索/ 新/ 机制/ 。
- 词性标注
  - 词性标注(Part-of-speech Tagging, POS)是给句子中每个词一个词性类别的任务。 这里的词性类别可能是名词、动词、形容词或其他。 下面的句子是一个词性标注的例子。
  - 其中，v代表动词、n代表名词、c代表连词、d代表副词、wp代表标点符号。
    - 词性标注结果：国务院/ni 总理/n 李克强/nh 调研/v 上海/ns 外高桥/ns 时/n 提出/v ，/wp 支持/v 上海/ns 积极/a 探索/v 新/a 机制/n 。/wp

语言技术平台（LTP）

- 依存句法分析：
  - 依存语法 (Dependency Parsing, DP) 通过分析语言单位内成分之间的依存关系揭示其句法结构。
  - 直观来讲，依存句法分析识别句子中的“主谓宾”、“定状补”这些语法成分，并分析各成分之间的关 系。
  - 仍然是上面的例子，其分析结果为：



代表性的工作

代表性的工作：2002年Pang等人用SVM方法在影评上进行情感分类

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86. Association for Computational Linguistics.

Thumbs up? Sentiment Classification using Machine Learning Techniques

Bo Pang and **Lillian Lee**
Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
{pabo,llee}@cs.cornell.edu

Shivakumar Vaithyanathan
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120 USA
shiv@almaden.ibm.com

Abstract

We consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods we employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. We conclude by examining factors that make the sentiment classification problem more challenging.

采用的方法

- 使用的机器学习模型：
  - 朴素贝叶斯（**Naïve Bayes, NB**）：
    - 基本原理：基于贝叶斯定理与特征条件独立假设的分类方法。
    - 给定训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入**x**，利用贝叶斯定理求出后验概率最大的输出**y**

$$P_{\mathrm{NB}}(c \mid d) := \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)}.$$

  - 最大熵（**Maximum Entropy, ME**）
    - 基本原理：学习概率模型时，在所有可能的概率模型（分布）中，熵最大的模型是最好的模型。

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)$$

- 支持向量机（**Support Vector Machines, SVMs**）
  - 基本原理：定义在特征空间上的间隔最大的线性分类器，支持向量机包括核技巧，使它成为实质上的非线性分类器。
  - 支持向量机的学习策略就是间隔最大化，可形式化为求解凸二次规划的问题。
  - 支持向量机的学习算法就是求解凸二次规划的最优化算法。

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \quad \alpha_j \geq 0,$$

实验结果

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

• 3.1.2 使用自定义打分函数的情感分类

- **Dave（2003）打分函数**
  - ✓第一步：用下面的等式为训练集中的每个词进行打分
    - $\text{score}(t_i) = \frac{Pr(t_i|C) - Pr(t_i|C')}{Pr(t_i|C) + Pr(t_i|C')}$
  - ✓第二步：将一个新文档$d_i = t_1 \dots t_n$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：
    - $class(d_i) = \begin{cases} C & eval(d_i) > 0 \\ C' & \text{其他} \end{cases}$
    - 这里的$eval(d_i) = \sum_j score(t_j)$
- 实验结果：在**7**种产品**13000**多条评论的数据集上，选用**bigram**和**trigram**特征能够达到最高的精确率**84.6%-88.3%**

3.2 基于无监督的情感分类

- **3.2.1 使用句法模板和网页检索的情感分类**
  - **Turney（2002）：将每个句法模板看做一个带约束的词性标签序列。**

算法包括三步：

✓ 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。

	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

✓ 用互信息PMI来估计所抽取短语的情感倾向性（SO）

$$\mathrm{PMI}(word_1, word_2) = \log_2 \left[\frac{p(word_1 \ \& \ word_2)}{p(word_1)\, p(word_2)} \right]$$

✓ 给定一个评论计算所有短语的SO（Semantic Orientation）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。

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

3.2.2 使用情感词典的情感分类

- 情感词典（观点词典）：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
  - 对于表达了正面/负面情感的文本表达（词或短语）都赋予了一个正的/负的SO值，最后将文档中所有情感表达的SO值求和。
  - 和为正，判定为正面情感；和为负，判定为负面情感；和为0，判定为中性情感。
- 情感转置词 （sentiment shifter）：否定词not和never，将SO值取反。
- 情感加强词和减弱词：
  - *This movie is very good.*
  - *This movie is barely any good.*

3.3 情感评分预测

- 情感打分：回归 或 分类方法
  - **Pang (2005): SVM**回归、基于一对多策略的**SVM**多分类（**OVA**）、度量标注的元学习
    - 结论表明：分类的效果要差
  - **Goldberg(2006):** 基于图的半监督学习问题
    - 每个节点表示一个评论文档，两个节点之间连接的权重是两个文档的相似度。
    - 相似度越高，表明两个文档的评分越接近。
  - **Qiu(2010):**文档的观点袋（**bag-of-opinion**），捕获观点中**n-gram**的情感强度。
    - 观点可以看做一个三元组（情感词，修饰词，否定词）
      - **Not very good : 情感词是good, 修饰词是very, 否定词是not**
    - 方法：有约束的岭回归（**ridge regression**）

3.4 跨领域情感分类

- 情感分类对训练数据所属领域非常敏感
  - 需要领域适应（domain adaptation）或迁移学习(transfer learning)的技术
    - 有标注的原始领域称为源领域，待测试的新领域称为目标领域
- 已有研究基于两个前提条件：
  - 1、需要新领域也有少量标注语料
    - Aue and Gamon 2005 提出的情感迁移分类算法：
      - （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；（SVM分类器）
      - （2）训练分类器同（1），仅适用目标领域中存在的特征；（SVM分类器）
      - （3）使用多领域分类器的集成，并在目标领域测试；（SVM分类器）
      - （4）结合使用目标领域的少量标注数据和大量未标注数据。（EM的半监督学习）
      - 注：4种方法中（4）效果最好

- 2、不需要新领域的标注数据
  - Blitzer等（2007）提出的结构化对应学习的方法（Structural Correspondence Learning, SCL）
  - （1）给定源领域的标注语料和未标注语料，目标领域的未标注语料，选择两个领域都频繁出现的m个特征，对源领域预测效果最好。
    - 这些特征称为 中轴特征
  - （2）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W。
    - 相关数为正表示在该领域的非中轴特征与中轴特征正相关，建立起两个领域的特征对应关系。

3.5 跨语言情感分类

- 任务：对多种语言的观点文档进行情感分类。
- 研究动机：
  - （1）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
    - 利用现有的机器翻译和英语情感分析系统的资源和工具
  - （2）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。
- 已有研究的典型方法：
  - 方法1：Wan（2008）利用英文的情感资源做中文的评论分类
    - 具体步骤：（1）用多个翻译引擎把每条中文评论都翻译为英文，等到多个英语版本；（2）用基于词典的情感分类方法对每份英文翻译进行分类；（3）通过情感打分获取评论的情感分类结果；（4）如果有中文情感词典，也可对中文情感评论进行分类，再和英文翻译结果进行结合。

3.5 跨语言情感分类

- 方法2：Wei and Pal (2010)
  - 基于迁移学习的方法进行跨语言情感分类
- 方法3：Blitzer et al（2007）
  - 基于SCL方法，选出由两种语言所共有的核心特征，为了减少数据稀疏性，他们向搜索引擎提交查询，试图找到那些和核心特征相关特征，然后用这些特征构建伪样例用以训练分类器。
- 方法4：Duh （2011）
  - 该文献认为领域不匹配不是由于机器翻译的错误引起。就算机器翻译结果完全正确也会造成跨语言分类准确率的下降，他们认为跨语言的自适应问题和单语言的自适应问题有本质不同，因此应该考虑不同的自适应算法。

3.6 文档的情绪分类

- 情绪分类的任务难度更大，原因分析：
  - （1）情绪的类别更多，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，不容易区分
- 已有主流方法：有监督学习方法
  - **Mishne and de Rijke(2006):** 针对博客数据进行情绪分类研究。利用博客上的心情标签进行有监督学习。特征主要选择代表每种情绪的词（**n-gram**）
  - **Lin （2007）**：利用**Yahoo**提供的中文新闻文章进行情绪识别。在新闻网页上，读者可以基于自己感受到的情绪对其进行投票。
    - 四种特征集：（1）基于汉字字符的**bigram**特征；
    - （2）包括中文分词后产生的所有词；
    - （3）文章的元数据，如：新闻记者、新闻分类、新闻事件地点、发布时间等；
    - （4）词的情绪类别，主要从情绪词典中获得。
  - 情绪词典：英文**wordnet**；中文：大连理工大学等

本章小结

- 文档级情感分类：针对整篇文档的整体观点和情感
- 存在的问题：
  - （1）不考虑情感或观点的评价对象。
    - 如果一条评论中评价多个实体，就不能得到更准确的评价信息，需要更细粒度的分析。
  - （2）文档级的情感分析不能提取细节，消费型用户不能从中了解到具体信息。