第 7 节 情感分析：什么是情感分析？

From Languages to Information CS124
——Sentiment Analysis https://web.stanford.edu/class/cs124/lec/sentimentvideoslides2019.pdf

积极或消极的电影反馈？

- Unbelievably disappointing.
- Full of zany characters and richly applied satire, and some great plot twists.
- This is the greatest screwball comedy ever filmed.
- It was pathetic. The worst part about it was the boxing scenes.

Google Product Search

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, **$100** nearby ★★★★☆ 377 reviews
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

1 star	2	3	4 stars	5 stars

What people are saying

ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints came out with great quality."

Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

Product summary | Find best price | **Customer reviews** | Specifications | Related items

$121.53 - $242.39 (14 stores)

☐ Compare

Average rating ★★★½☆ (144)

★★★★★		(55)
★★★★☆		(54)
★★★☆☆		(10)
★★☆☆☆		(6)
★☆☆☆☆		(23)
☆☆☆☆☆		(0)

Most mentioned

Performance		(57)
Ease of Use		(43)
Print Speed		(39)
Connectivity		(31)

More ▼

Show reviews by source

- Best Buy (140)
- CNET (5)
- Amazon.com (3)

情感分析（Sentiment analysis）的其他说法

- 观点抽取（Opinion extraction）
- 观点挖掘（Opinion mining）
- 情感挖掘（Sentiment mining）
- 主体性分析（Subjectivity analysis）

为什么要进行情感分析？

- 电影：评论是正面的还是负面的？
- 产品：人们认为新一代iPhone怎么样？
- 公众情绪：消费者信心如何？绝望增加了吗？
- 政治：人们对这个候选人或这个问题的看法是什么？
- 预测：从情感预测选举结果或市场趋势

谢勒类型学（Scherer Typology）中的情感状态

- 情绪（Emotion）：由一定原因引发的同步反应
  - angry, sad, joyful, fearful, ashamed, proud
- 心情（Mood）：没有明显原因引发的长期低强度的主观感受变化
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- 人际关系立场（Interpersonal stances）：在特定互动中对一个人的情感立场
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **态度（Attitudes）：对特定的人或事物的持久的、带有主观色彩的偏好或倾向**
  - **liking, loving, hating, valuing, desiring**
- 人格特质（Personality traits）：稳定的性格倾向和典型的人格倾向
  - nervous, anxious, reckless, morose, hostile, jealous

情感分析

- 情感分析是对态度（Attitudes）的检测

"对特定的人或事物的持久的、带有主观色彩的偏好或倾向"

1. 持有人（来源）的态度
2. 目标（方面）的态度
3. 态度的类型
   - 一系列的类型
     - ***Like, love, hate, value, desire,*** **etc.**
   - 或者（更为常见的）简单的对积极性的加权：
     - ***Positive, negative, neutral,*** **together with** ***strength***
4. 态度的范围
   - 句子或整个文档

情感分析

- 简单任务
  - 这篇文章的态度是积极的还是消极的？
- 更为复杂的
  - 将文本中的态度按1-5的级别进行排序
- 高级的
  - 检测目标、来源，或更复杂的态度类型

情感分析

- 简单任务
  - **这篇文章的态度是积极的还是消极的？**
- 更为复杂的
  - 将文本中的态度按1-5的级别进行排序
- 高级的
  - 检测目标、来源，或更复杂的态度类型

第6节 情感分析：基线算法（A Baseline Algorithm）

学习Natural Language Processing with Deep Learning CS224N/Ling284

—— Lecture 2: Word Vectors https://web.stanford.edu/class/cs224n/lectures/cs224n-2017-lecture2.pdf

Lecture Notes 1 https://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes1.pdf

电影评论中的情感分类

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

- 极性检测：
  - IMDB电影评论是正面的还是负面的？
- 数据：*Polarity Data 2.0:*
  - http://www.cs.cornell.edu/people/pabo/movie-review-data

IMDB data in the Pang and Lee database

✓

when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .

cool .

october sky offers a much simpler image–that of a single white dot , traveling horizontally across the night sky . [. . .]

✗

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

it’s not just because this is a brian depalma film , and since he’s a great director and one who’s films are always greeted with at least some fanfare .

and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

基线算法（Baseline Algorithm，改编自Pang and Lee）

- 符号化（Tokenization）
- 特征提取（Feature Extraction）
- 用不同的分类器进行分类
  - Naïve Bayes
  - MaxEnt
  - SVM

情感标记问题

- 处理HTML和XML标记
- Twitter标记（名字，#）
- 字母大小写（中文不存在）
- 电话号码，日期
- 表情
- 有用的代码：
  - Christopher Potts sentiment tokenizer
  - Brendan O’ Connor twitter tokenizer

提取情感分类的特征

- 如何处理否定
  - I **didn't** like this movie

    vs
  - I really like this movie
- 用哪些词进行分析？
  - 仅用形容词
  - 所有词汇
    - 所有词的效果显示都很好，至少在这些数据上

否定

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).
Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

从出现否定到最近的标点符号之间，在每个单词中添加“NOT_”

didn’t like this movie , but I

didn’t NOT_like NOT_this NOT_movie but I

Naïve Bayes：朴素贝叶斯

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

- 左边代表的是这个评论的极性类型，就是右边概率最大的情况下的极性类型
- $P(c_j)$指该极性类型出现的概率
- $\prod_{i \epsilon positions} P(w_i, c_j)$指在当前极性条件下文档中各个词出现的概率的乘积

Naïve Bayes：朴素贝叶斯

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

- 这里用Laplace(+1 smoothing)转换
- 其中 V 就是当前训练文本的词汇量

二值化（布尔特征）多项式朴素贝叶斯

- Intuition:
  - 对于情感领域（可能对于其他文本分类域）
  - 单词出现（word occurence）可能比单词频率（word frequency）更重要
    - *“fantastic”*的出现提供了很多信息
    - 某些单词即便出现5次也不会提供很多信息，例如“*we*”
  - 布尔多项式朴素贝叶斯
    - 将每个文档中的所有出现的词的频率统计为1

布尔多项式朴素贝叶斯：模型训练

- 从训练语料中，提取词汇表（*Vocabulary*）
- 计算$P(c_j)$项

For each c_j in C do

$docs_j \leftarrow$ all docs with class $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

布尔多项式朴素贝叶斯的特点为第一部去重，然后把所有文档合并为一个大文档；α在这里为Laplace处理。

- 计算$P(w_k | c_j)$

- Remove duplicates in each doc:
  - For each word type w in doc_j
    - Retain only a single instance of w
- $Text_j \leftarrow$ single doc containing all $docs_j$
- For each word w_k in *Vocabulary*

  $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

对于验证集

- 从验证集中取出所有重复单词
- 用相同的公式计算 NB：

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

一般贝叶斯 VS 布尔贝叶斯

- 一般贝叶斯

Normal	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- 在测试集上，C类Chinese出现频次为 5，P(Chinese|C) = 5/8；
- 在验证集上的频次为 3

一般贝叶斯 VS 布尔贝叶斯

- 布尔贝叶斯

Boolean	Doc	Words	Class
Training	1	Chinese Beijing	c
	2	Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Tokyo Japan	?

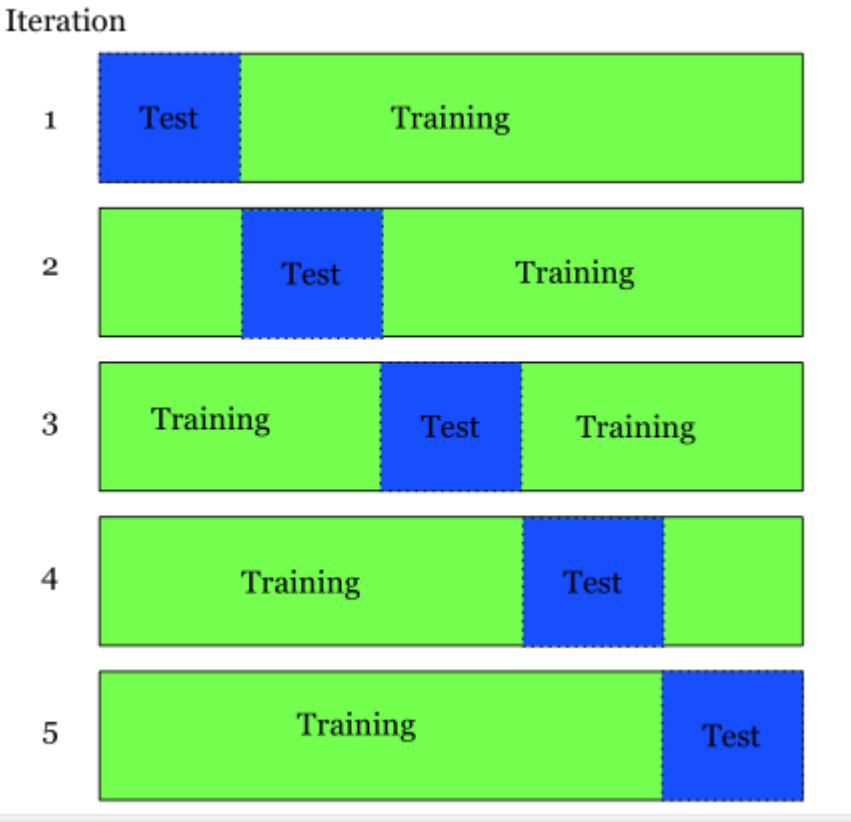
- 在测试集上，C类Chinese出现频次为 3，P(Chinese|C) = 3/6；
- 在验证集上的频次为 1

二值化多项式朴素贝叶斯

- 经过测试，Boolean naïve Bayes 的效果比计数所有词的一般 naïve Bayes 要好
  - Boolean naïve Bayes 与 multivariate Bernoulli naïve Bayes（MBNB）不同，后者不适用于情感分析相关的问题
- 频次统计出了所有词或去重只保留一个的方法外，还可以利用$log(freq(w))$求解，该词频介于上述两种方法之间，可能取得更好的效果

交叉验证（Cross-Validation）

- 将数据分为10份
  - 每份里需要包含相同数量的正面评价和负面评价吗？
- 对于每份文件夹
  - 选择一份文件夹为临时测试集
  - 对其余 9 份进行训练，在测试集上计算性能
- 报告10次运行的平均性能



注：一般来说，除此以外最后会用一个没有在交叉验证数据集中的验证数据集做最后测试，以防止过拟合

分类的其他问题

- MaxEnt 与 SVM 分类器一般在大量数据下的表现比 naïve Bayes 要好

顺序效应（Florida Expectation Problem）

- 指前面描述了很长的表示期待的语句，仅仅为了对比烘托出后文的不满情绪
  - “This film should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a good performans. However, **it can’t hold up**.”

第6节 情感分析：情感词典（Sentiment Lexicons）

学习Natural Language Processing with Deep Learning CS224N/Ling284

—— Lecture 2: Word Vectors https://web.stanford.edu/class/cs224n/lectures/cs224n-2017-lecture2.pdf

Lecture Notes 1 https://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes1.pdf

The General Inquirer

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

- Home page: http://www.wjh.harvard.edu/~inquirer
- List of Categories: http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Spreadsheet: http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls
- Catrgories:
  - Positiv (1915 words) and Negativ (2291 words)
  - Strong vs Weak, Active vs Passive, Overstated vs Understated
  - Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
- Free for research use

极性词典之间的差别

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
General Inquirer			520/2306 (23%)	1/204 (0.5%)
SentiWordNet				174/694 (25%)
LIWC				

由表中数据可见，各词典对于极性的判断基本上比较一致

分析IMDB的评论极性

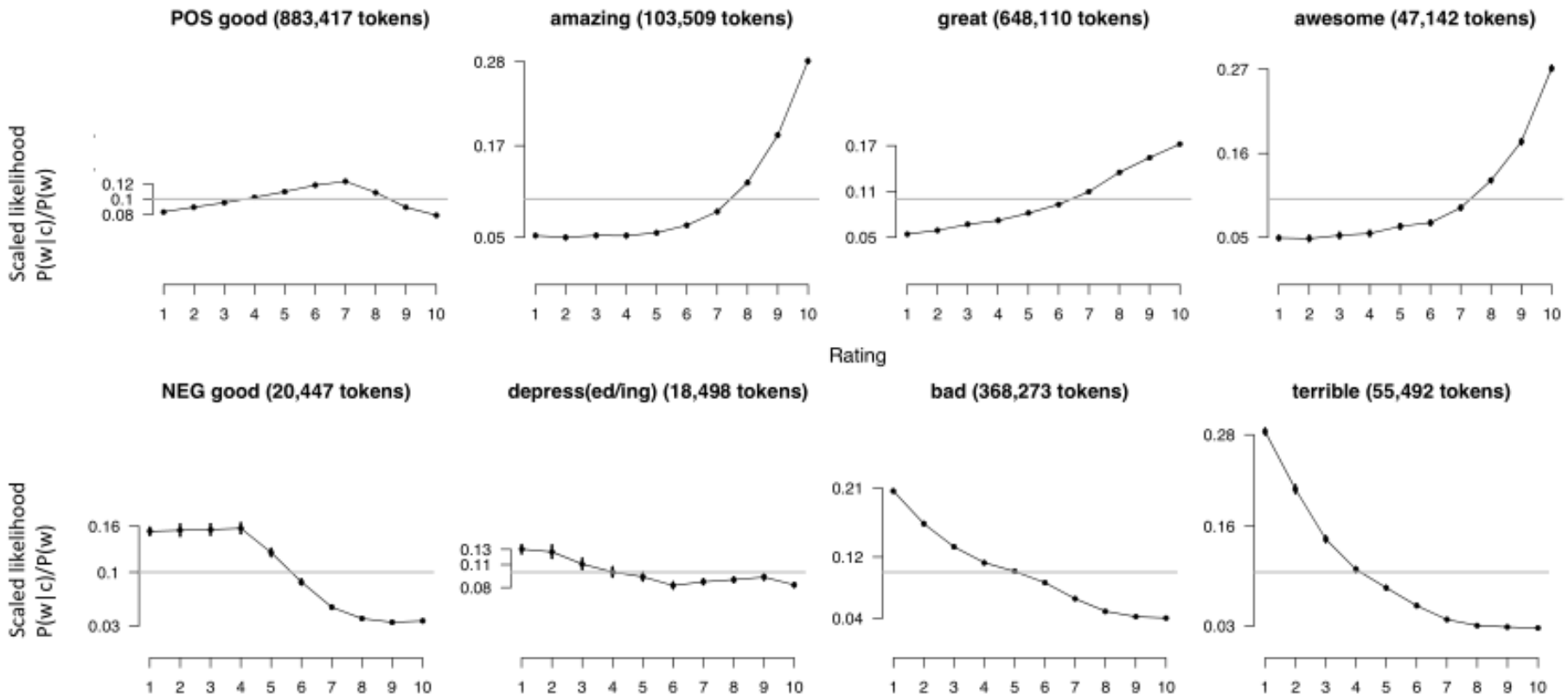
- 将1-3星计数为差评类，但不要使用原始的差评类中的词的个数用于极性分析，因为虽然评高分的人比评一分的人少，但是留下评论的更多的是好评的人（满意的人倾向于留下评论），因此直接使用会有偏差，需要用似然度（likelihood）来衡量：

$$P(w|c) = \frac{f(w,c)}{\sum_{w \epsilon c} f(w,c)}$$

- 其中$f(w,c)$指 c 类中 w 的个数，分母为该类所有词的个数
- 然后归一化，使词与词之间的频率可以相互比较：

$$\frac{P(w|c)}{P(w)}$$

一些词在各评分中出现的频率



其他的情绪特征：逻辑否定（Logical Negation）

- 逻辑否定（no, not）是否与负面情绪相关？
- Potts experiment:
  - 否定词常出现于负面情绪

第6节 情感分析：其他情感分析任务

学习Natural Language Processing with Deep Learning CS224N/Ling284

—— Lecture 2: Word Vectors https://web.stanford.edu/class/cs224n/lectures/cs224n-2017-lecture2.pdf

Lecture Notes 1 https://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes1.pdf

单句情感

- 单句情感：寻找情感描述的对象

- The food was great but the service was awful

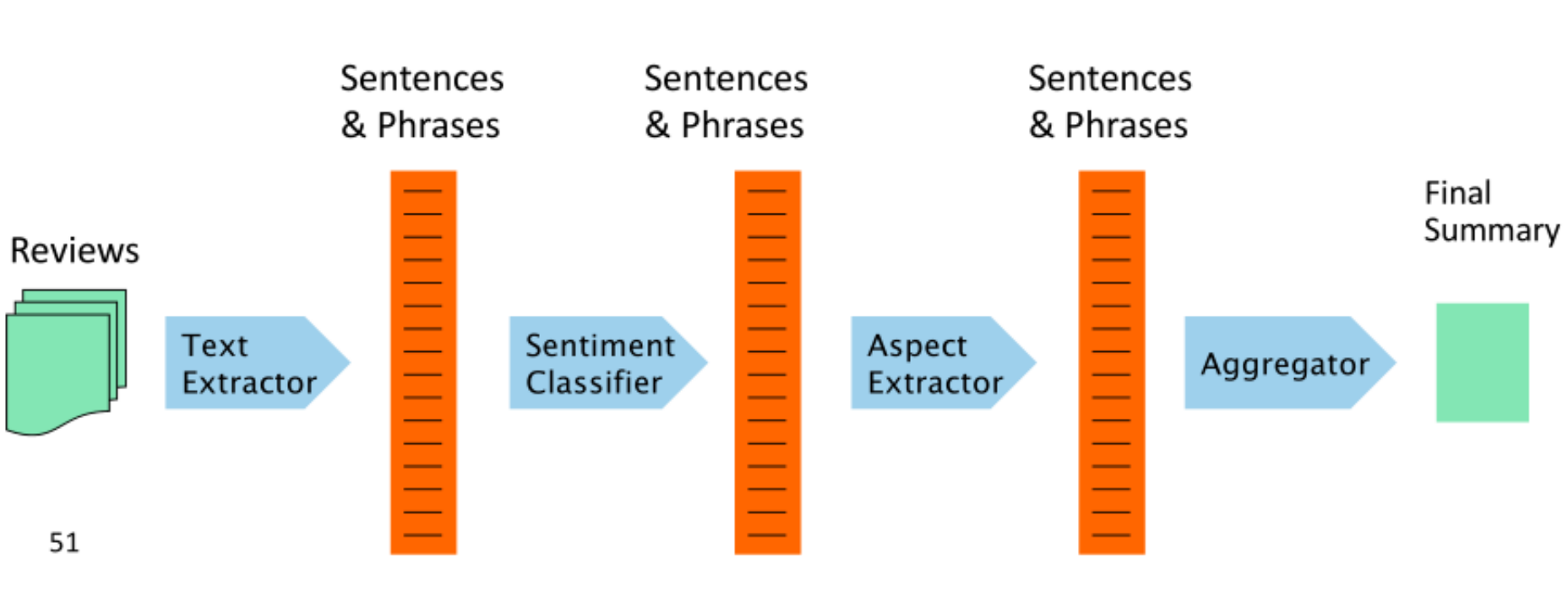
情感分析中寻找对象/属性/目标

- 高频短语+规则：首先获取评论中出现的高频短语，然后制定规则，比如对评价食物，我们可以将出现在情感词之后的高频词作为情感的对象——高频词为fish tacos，那么评论中出现great fish tacos时，我们就认为fish tacos是great的情感描述对象

情感分析中寻找对象/属性/目标

- 对于描述内容的方向比较明确的评论，可以预先定义对象，然后用监督学习的方法进行对象分类。比如对于餐馆评论而言，一般描述的对象为food，service，value等。首先将一些餐馆评论语句人工打标为上述标签，然后作为训练集训练一个描述对象的分类器。

总体流程



分句 -> 情感极性分类 -> 提取主题 -> 汇总得到增提评论

基线方法假设类具有相同的频率！

- 样本不平衡问题
  - 不同极性的评论数量差距太大（例如：10^6好评与10^4差评），会导致分类器模型参数一场。解决方法为重抽样（使好评数与差评数均衡）或者采用代价敏感学习（cost-sensitive learning），比如在训练SVM分类器的时候，将稀有样本错误分类的惩罚加大。
- 处理打分问题（例如5星）
  - 可将其转化为二元分类问题，比如小于2.5星的视为负面评价，大于2.5星的视为正面评价
  - 直接将星与极性强度用线性回归或其他方式拟合

总结

- 情感分析本质上还是分类问题（二分类或者回归）
- 否定词在情感描述中很重要
- 使用所有词的naïve Byes模型在一些问题上表现较好，而使用子集短语模型则在另一些上表现较好
- 构建极性词库流程：手动标记种子极性词库，然后用半监督学习方式扩展词库
- 除了态度（attitudes）以外，Scherer情感状态的其他类型，都是情感分析的重要方向

Thank you !