基于情感词典和规则

特征工程+ 机器学习

深度学习

（）：括号内 内容为 添加到PPT相应页注释的内容

**基于情感词典和规则**

-------------------------

一般步骤：

a.分词

b.设置情感词典：情感词典一般包括5个词典，即正面情感词典、负面情感词典、否定词典、程度副词词典和行业情感词典

（1.程度副词 比如“非常”、“极其”等，会对情感有增强作用，最终影响整个短语或句子的总体情感倾向

2.行业情感词典，即只在某些特定领域具有情感倾向的词，比如对于汽车，“塑料感”、“底盘硬”其实是表达负面情感

）

c.基于规则匹配

（基于分词和情感词典，即可以根据人们平时的语言表达习惯设置一些规则来计算文本的情感倾向，比如每遇到一个正面情感词则+1分，遇到负面情感词则-1分，遇到否定词则乘以-1将情感反转，遇到程度副词则将情感分数乘以一个放大系数。最后根据计算出的分数判断情感倾向，分数为正数则判断为正面情感，负数则判定为负面情感，正负相抵则判定为中性

）

##

**优点**：不需要标注好的训练数据

##

**不足**：可扩展性非常差，需要人工一个个配词典加规则，才能识别足够多的情感倾向

（并且需要完全精确匹配，因此召回率比较差。同一个含义人可以有很多种表述方式，尤其在情感分析领域，常常出现双重否定等复杂句式，规则必须设计得足够复杂才能进行识别。此外，当配的规则比较多的时候，不同规则往往会发生互相冲突，这时候情感分析结果就不可控了，会导致准确率下降）

**特征工程+ 机器学习**

--------------

a. 特征提取：

（文本挖掘领域最常用的也是最简单的特征是词袋模型（bag of words），即将文本转换为基于词语的一个向量，向量的每一维度是一个词语，词语可以基于分词得到，也可以基于N-Gram模型得到。每一维度的特征取值也可以有多种计算法方式，比如经典的one-hot编码和tf-idf值）

b. 模型训练：基于文本的标注类别和提取好的特征向量，即可以使用机器学习算法进行训练，

（模型训练完成之后即可用于判别文本的情感倾向）

##**优点**：相比较于基于规则的方法，只需要准备好标注数据，设计好特征提取方法，模型可以自动从数据中学习出一个复杂的高维分类模型实现情感分析

##**缺点**：其效果主要取决于特征工程，即提取的特征是否能足够很好的区别正面和负面情感。

（在相同的特征下，如果只使用简单分类器，那选择不同的分类算法，效果差别不会太大。要做好特征工程，非常依赖于人的先验知识，即需要我们对数据进行足够深入的观察和分析，把那些对区分正负面情感最有用的特征一个一个找出来

）

**重要**：**特征工程做深入了也需要依赖情感词典和规则方法，但不是直接判定文本的情感倾向，而是将规则命中的结果作为一维或者多维特征，以一种更为“柔性”的方法融合到情感分析中，扩充我们的词袋模型**\*\*

**深度学习**

---------------------

1. 词语转成词向量：

（google的word2vec算法是目前应用最广泛的词向量生成算法，实践证明其效果是非常可靠的，尤其是在衡量两个词语的相似度方面。Word2vec算法包含了CBOW（Continuous Bag-of-Word）模型和Skip-gram（Continuous Skip-gram）模型。简单而言，CBOW模型的作用是已知当前词Wt的上下文环境（Wt-2，Wt-1，Wt+1，Wt+2）来预测当前词，Skip-gram模型的作用是根据当前词Wt来预测上下文（Wt-2，Wt-1，Wt+1，Wt+2）。因此，一次词向量事实上是基于词语的上下文来生成的，也就具备了词袋模型所不具备的表意能力。）

1. 利用深度学习框架进行训练：目前比较广泛使用的是LSTM

（词转成固定维度的词向量之后，一个文本也就自然而然可以形成一个矩阵。以矩阵作为输入的深度学习算法，第一个想到的自然是在图像识别领域获得过成功的卷积神经网络（CNN）。但CNN在文本挖掘领域的运用具有一定局限性，因其每层内部的节点之间是没有连接的，即又丢失了词与词之间的联系。前面已经多次强调，词语的上下文关系对文本挖掘是至关重要的，尤其对情感分析，情感词（“喜欢”）和否定词（“不”）、程度词（“很”）的搭配会对情感倾向产生根本性的影响。）因此目前比较广泛使用的是LSTM（Long Short-Term Memory，长短时记忆），LSTM能够“记住”较长距离范围内的上下文对当前节点的影响）

\*\***优化问题**---》词袋模型 词语之间的语义是相互独立 丢失了文本的上下文信息 ---》改进 词嵌入（word-embeding）

（普遍应用的词袋模型隐含了一个假设，即词语之间的语义是相互独立的，因而丢失了文本的上下文信息。但真实情况往往并非如此，同一个词语在不同的语义环境下是可以具有不同语义的。词袋模型还会导致向量空间特别大，一般都是数十万维。对于评论这种短文本，转换成的向量会特别稀疏，也造成了模型的不稳定性\*\*）

##相比于传统机器学习方法，深度学习至少有**3大直接优势**：

##

1）**无需特征工程**：（深度学习可以自动从数据中学习出特征和模型参数，省去了大量繁杂的特征工程工作，对行业先验知识的依赖也降低到最小程度。）

##

2）**考虑语义上下文**：

（深度学习在处理文本数据的时候，往往是先把词语转成词向量再进行计算，词向量的生成考虑了一个词语的语义上下文信息，也就解决了词袋模型的局限性。）

##

3）**大幅减少输入特征维度**：

（由于使用了词向量，特征维度大幅减少，可以降低到百的量级，同时也使得文本向量变得“稠密”，模型变得更加稳定。）

##优点：：：\*基于深度学习的文本情感分析，相比传统机器学习，效果可以提升15%左右，而且省去了繁复的特征工程工作，将人工依赖降低到最低程度\*

参考资料：https://zhuanlan.zhihu.com/p/27068121

=================================================================

情感分析的主要任务

\*1 情感信息抽取

2 情感分类

3 情感检索与归纳\*

1 情感信息抽取包括：

a 识别情感表达者

b 评价对象

c 情感观点

2 情感分类

-----------

##

a 基于词典和规则的情感分类算法

情感词典 能够体现文本的非结构特征

\*\*基于情感词典的情绪分类方法

》》基于句法分析与情感词典相结合的方法

：：：利用情感词典从广告上下文中识别感情句，根据主题和关键字提取用户的态度

》》扩充情感词典特征和主题相关特征

》》中文情感分析， 机器翻译将中文商品评论翻译成英文评论，然后在进行情感分析的方法

》》基于互信息的情感分类算法

：：：提取文本磁性的基础上，根据预定义的规则选取形容词、副词的搭配，对文本所有搭配的互信息差求和，判断情感分类

？解决问题： 情感词在不同的领域或语境中有不同的情感极性

》》基于“主题-句子”关系的情感分类方法

：：：在情感词上同时标记主题和情感2种标签，并利用句子的主题标签采样代替词的主题标签采样，所辖了词与词之间的主题联系\*\*

##

b 基于机器学习的情感分类算法

\*\*

1） 有监督的机器学习算法

》》通过对比一元特征、二元特征、形容词打分、位置等多个特征和特征权值选择策略。并着重比较了SVM、朴素贝叶斯和最大熵算法的分类效果

》》基于自适应递归神经网络的情感分类算法

：：：通过上下文和句法规则对词的情感标记进行自适应传播，实现目标依赖的情感分类

》》基于消息级微博情感分析的深度学习系统

：：：将特定情感词向量（sentiment-speciafic word embedding）与手工选择的表情符号、语义词典等特征相结合，并利用SVM进行情感分类

2） 半监督学习的情感分类算法

》》基于半监督特征提取的情感分类系统

：：：系统融合谱聚类、主动学习、迁移学习等不同方法提取情绪特征，应用迁移学习的方法完成整个情感分析系统的构建

》》提出一种递归自编码半监督学习的情感分析模型，

：：：模型在构建短语向量表示时，可以很大程度的保留情感信息，提高了预测情感的准确率

》》通过将先验知识嵌入到学习结构中，提出一种基于模糊深度置信网络的半监督学习情感分类方法

：：：该方法继承深度置信网络强大的抽象能力，还具备对情感数据的模糊分类能力

========================

依据情感分析的粗细粒度

可以分为 ：篇章级情感分析 和 句子级情感分析

## 篇章级和句子级情感分析的前提假设是整个文本和句子只包含一种情感

句子级情感分析

\*步骤： 1 主客观分类

2 将主观语句中的情感倾向进行分类-》正负情感

\*\*

\*\*aaa 粗粒度到更细的粒度？？？ 随着各种评论文本的变化，有这个需求

>e.g 产品评论文本的整体情感倾向并不一定能和产品各个属性的情感倾向保持一致。评价对象中的多个属性时，不能满足要求。在这种情况下，需要识别属性词、情感词及其之间的关系

>>e.g ，例如手机评论中，手机作为评论对象包含了电池、手机屏幕、照相功能和手机操作系统等属性。评论发布者往往会从评论对象属性层面对手机的各个功能部分做出评价

这种识别属性机器对应的情感倾向需要进行细粒度的情感分析，也被称为基于属性的观点挖掘

细粒度中的2个问题：：

1 隐式情感 “Q5 太贵了啊，买不起” 隐藏了 价格

2 客观句中的情感表达 “the computer crashed every day”

细粒度情感分析步骤：

1 对评论对象属性及其对应情感词进行识别

2 对识别出的对象属性的情感进行极性分类

3 对分类结果进行汇总

其中的关键---》评价对象属性及其对应的情感词 抽取

细粒度情感分析的难点出现在：

1 隐氏情感分析， 现阶段对隐氏情感分析主要集中在 隐氏评论对象抽取， 但对 客观文本中隐藏得情感倾向识别研究较少

2 跨领域情感分析， 主要是不同领域间的情感表达差别较大