098 | 什么是文档情感分类?

2018-05-18 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:19 大小 2.89M

到目前为止，我们讲完了对话系统的基础知识。一般来说，对话系统分为“任务型”和“非任务型”这两种基本类型。针对任务型对话系统，我们重点介绍了其各个组件的任务，以及这些组件都有哪些模型给予支撑。针对非任务型对话系统，也就是“聊天机器人”，我们主要介绍了如何利用深度学习技术来对一个聊天机器人进行建模，以及非任务型对话系统所面临的挑战都有哪些。

今天，我们转入文本分析的另外一个领域，同时也是在实际系统中经常会使用的一个子领域，那就是**文本“情感分析”**（Sentiment Analysis）。所谓情感分析，就是指我们要针对一段文本来判断这段文本的文字“色彩”，到底是褒义，还是贬义，到底是抒发了什么情感。

文本情感分析是一个非常实用的工具，比如，我们需要分析用户对于商品的评价带有什么样的情感，从而能够更好地为商品的推荐和搜索结果服务。再比如，通过文本的情感分析，我们可以了解到用户针对某一个时事的观点异同，以及观点分歧在什么地方，从而能够更加清晰地了解新闻的舆情动态。

今天，我们首先从最基础的**文档情感分类**（Document Sentiment Classification）这个问题说起。

基于监督学习的文档情感分类

文档情感分类属于文本情感分析中最基本的一种任务。这种任务的假设是，一段文本的作者通过这段文本是想对某一个“**实体**”（Entity）表达一种情绪。这里的实体其实包括很多种类型的对象，比如可能是商品，某个事件，也可能是某个人物。我们这里讨论的文本单元可以是一个文档，也可以是一个句子等其他的文本段落。

值得注意的是，我们在这一类任务中，限制一个文本单元只表达，或者主要表达一种情感。很明显，这种假设是比较局限的。一般来说，在实际的应用中，一个文本单元，特别是比较长的单元例如文章，则往往包含多于一种的情绪。因此，我们可以看到**文档情感分类其实是一种简化了的情感分析任务**。

同时，一个文本单元还可能对多个“实体”进行情感表达。比如一个用户针对某种款式相机的多个方面进行了评价，那么每一个方面都可以作为一个实体，而这种时候，用户的情感可能就更难仅以一种情感来加以概括了。

在最基本的文档情感分类的情况下，我们往往把这类任务转化成为一种**监督学习任务**，也就是说，我们希望通过一个有标签的训练集学习到一个分类器（Classifier）或者回归模型（Regression），从而能够在未知的数据上预测用户的情感。

这里往往有两种形式的监督学习任务。一种是把文档分类为几种，最简单的情况下是两种情感。这就是**二分或者多类分类问题**。另外一种则是认为文档会有一种情感，但是每一种情感之间有好坏的顺序区分，比如，评分“好”，就比“一般”要好，也就是说，这些评分之间有一个次序问题。那么，很多时候，这种问题会被归结为一种**“次序回归”（Ordinal Regression）问题**。

在明确了我们需要构建什么样的监督学习任务以后，对于这些任务而言，如何选取“特性”（Feature）就是一个很重要的工作了。诚然，对于每一个具体的任务而言，我们往往

需要选取不同的特性，但是在过去的很多实践中，经过反复验证，有一些特性可能会有比较好的效果。我在这里做一个简单的总结。

首先，我们曾经多次提到过的“**词频**”（Term Frequency）以及更加复杂一些的**TF-IDF词权重法**都是经常使用的文字特性。在文档情感分类中，这一类特性被认为非常有效。

另外一种使用得比较频繁的特性就是“**词类**”（Part of Speech）。词类提供了句子中每个词的成分，比如哪些词是动词，哪些词是名词等等。这些词性可以跟某种特定的情感有很密切的联系。

还有一种很直观的特性就是“**情感词汇**”。比如，我们已经知道了“好”、“不错”等词表达了正向的情感，而“差”、“不好”、“不尽人意”等词表达了负向的情感。我们可以事先收集一个这类情感词汇的集合。这个集合里的词汇可以跟最后文档的情感有很直接的联系。

最后，需要指出的是，如何开发一个合适的特性往往是文档分类的重点工作。

除了特性以外，在文档情感分类这个任务中，传统上经常使用的文字分类器有“**朴素贝叶斯**”（Naïve Bayes）分类器、“**支持向量机**”（Support Vector Machines）等。

基于非监督学习的文档情感分类

情感词汇已经为我们对大段文字乃至整个文档的分类有了很强的指导意义，因此，也有一些方法寻求利用**非监督学习的方式**来对文档进行情感分类。注意，这里所谓的非监督学习，是指我们并不显式地学习一个分类器，也就是说，不存在一个训练数据集，不需要我们提前收集数据的标签。

这一类思想的核心其实就是设计一套“**打分机制**”（Scoring Heuristics），来对整个文档做一种粗浅的判断。当然，这种打分机制背后都有一种理论来支撑。

比如，有一种打分模式依靠首先识别的“词类”进行分析，特别是大量的相邻的两个词的词性，诸如“特别好”。这里，“特别”是副词，“好”是形容词，然后就可以得出在某些情况下，副词和形容词的这种搭配特别多的时候，并且在正向的情感词比较多的时候，整个文档也许就是比较偏向正向的情感。

我们需要指出的是，这种方法虽然听上去比较“山寨”，但是对于很多产品和项目来说，获取大量高质量的标签信息往往是非常耗时，甚至是不可能的，例如上百万的用户对产品的评价数据。因此，在没有训练数据的情况下，利用某种打分机制，可以通过最简单的一些情感词库开发出文档情感分类的算法，这其实也不失为一种**快速迭代的方式**。

总结

今天我为你介绍了一类基础的文字情感分析任务——文档情感分类的基本技术要点。

一起来回顾下要点：第一，我们讲了基于监督学习的文档情感分类任务以及这类任务下的重要特性和模型；第二，我们聊了如何在没有大规模训练数据的基础上进行非监督的文档情感分类。

最后，给你留一个思考题，如何把文档情感分类任务扩展到可以针对多种实体多种情感的分析呢？

欢迎你给我留言，和我一起讨论。

上一篇 [illegible]
下一篇 099 | 如何来提取情感“实体”和“方面”呢?

精选留言 (1)

写留言

paradox
2019-01-06

初学者，有个突然的想法：
可不可以先通过找出具有一些高信息量的词，从而得到这段文本不同的主题，然后把主题和文本放在一起在针对这个主题去预测情感，是不是就是多种实体多种情感的分析呢?