Chapter 9 情感分析（研究进展、现状&趋势）

什么是情感分析?

- 狭义：指利用计算机实现对文本数据的观点、情感、态度、情绪等的分析挖掘
- 广义：包括对图像视频、语音、文本等多模态信息的 情感计算
- 目标：建立一个有效的分析方法、模型和系统，对输入信息中某个对象分析其持有的情感信息，例如观点倾向、态度、主观观点或喜怒哀乐等情绪表达。

（文本）情感分析分类

情感资源构建

- 情感资源通常体现为一些带有情感倾向标注的词或短语，这些资源成为各种情感分析任务的重要资源支撑
- 类别体系的研究
  - 从情感倾向、情感表达强弱等方面对情感表达进行区分的类别体系，最常见的包括正、负倾向、主客观，以及细粒度的表达情感强度的强弱区分
- 不同粒度的情感资源研究
  - 从资源词条的文本粒度来说，有词汇级别、短语级别和属性级别，而往往更细的粒度需要的领域知识更多，难度更大
- 构建方法的研究
  - 手工构建、基于词典扩展和基于语料库构建的方法

情感信息的质量分析

- 对信息内容本身的判别，包括评论内容可信度分析（Credibility）、垃圾评论识别（Spam）评论内容的可用性（helpfulness）分析等
- 对信息内容提供者的判别，甄别虚假用户

情感分类（基础任务）

- 对给定的信息内容，依据情感类别体系进行分类（文本分类任务）或评级（序回归任务）
- 从输入文本的粒度来看，可以分为篇章级、句子级、短语级、对象和属性级
- 从所采用的方法来看，可以分成无监督学习、半监督学习、有监督学习方法
- 从任务的定义上，可以分成主客观分类，情感倾向极性分类，以及情感倾向强度评级（例如 1~5 分，或 1~10 分）。

情感信息抽取

- 情感信息抽取是情感分析中的细粒度任务，其核心的目标是抽取观点对象、评价表达、 对象和评价之间的搭配等
- 抽取观点对象：通常有关于观点持有人、观点所针对的目标、对象的细粒度属性等不同层次的情感识别与抽取
- 评价表达：通常是从输入内容中抽取情感词、情感表达式等内容，包括隐性表达（即通过事实类描述或其它隐晦描述）和显性表达（即具有明显的观点描述）
- 对象和评价之间的搭配：不仅要识别观点对象或属性及针对其的情感评价

多模态情感分析

- 传统的情感分析任务大多是在文本信息上进行的。多模态的情感分析是指从图像、视频、 语音、文字等多模态的数据中分析情感、情绪的表达。
- 单模态数据的情感分析，例如针对语音数据、面部视觉信息进行情感情绪识别
- 多模态融合的情感分析 例如从语音+视觉的数据中分析情绪表达，从图像+文字的数据中分析情感表达，从语音+文字的数据中分析观点表达等。

情感分析方法

规则为主的情感分析方法（早期）

- 利用一些已知的情感资源，并结合一些句法规则（如同、反义词，否定、转折、递进等），并结合一些统计量，从而进行情感资源构建或者情感分类操作
- 缺点：需要较多的资源（词汇资源、各种规则），并且规则总结和挖掘，不可避免的需要介入手工检查

传统机器学习的情感分析方法

- 特征：词性、情感词汇、句法依赖、 情感变换词（not，no，never，neither）等
- 近年来主题模型也成为情感抽取的一类重要方法。在这一类方法中观点对象和情感词都被当作是主题信息。一个主题中往往包含了数个概率较高的词，因此这类方法在抽取的同时也完成了词的聚类（基于 pLSA 的特征-情感混合模型）

基于深度学习的情感分析方法（目前几乎霸榜NLP tasks）

词向量的表示

- 在词向量的表示学习基础上，加入情感相关的目标函数，进行联合训练，以期望得到与情感信息相关的词向量表示
- 根据词性选择合成函数，以及学习一个词性的嵌入向量，根据子节点向量、词性向量合成父节点的向量（通常形容词扮演更重要的角色）

采用自动编码器进行文本的表示学习

- 1. 简单的编码器， 将文本的词袋表示（词表上的稀疏向量表示）转成隐藏层上的表示，学习的目标是最小化原始输入和重构表示之间（隐藏层表示经过非线性变换得到）的误差。
  应用场景：领域自适应、跨语言的表示或跨模态的数据表示
- 2. 面对情感分析任务，现有研究者已经把情感分类和领域分类的监督信息加入到优化目标函数中，使得所得到的表示具有一定的情感表达的特点。

面对句子级情感分析任务

- 种在句法成分树上进行递归编码的深度学习模型，通过在每个内节点上加入情感标注的监督信息，和重构误差一起进行优化，在句子级别的情感分类上较传统词袋模型获得了大幅提高
- CNN、LSTM等

发展趋势

面向社交媒体开放域文本的情感分析

- 难点：评论对象或属性更加难以抽取，表达更加隐晦，甚至不存在明显属性描述词；观点表达更加多样，许多话题不存在明显的观点评价词；理解情感表达需要更多的上下文，例如评论、转发、反讽中需要通过上下文才能对内容进行充分理解

基于上下文感知的情感分析

- 要求在理解当前内容时候，考虑各种形式的上下文
- 难点：1）基于上下文感知的情感资源构建方法；2）基于上下文相关的情感分类，包括篇章级、句子级、对象级、对象属性级、社交媒体的上下文。

跨领域跨语言情感分析

- 原因：情感语义计算极大依赖于情感资源（包括情感词典与标注语料），而情感资源又通常跟领域、语言密切相关。但是社交媒体上用户生成文本涉及众多的不同领域，以及不同的语种（例如中文、英文、日文，以及少数民族语言等）
- 亟待提出崭新的跨领域跨语言文本情感计算理论与方法，破除领域或语言壁垒。

基于深度学习的端到端情感分析

新的情感分析任务

- 情感解释： 挖掘与分析观点情感的原因。比如在社交媒体上，面对热门事件或开放性话题，如何分析群体情感的演变模式和原因分析。
- 反讽分析：反讽是社交媒体上一类特殊的语言现象，网民有时候会利用反讽来表达与文本字面相反的语义或情感倾向。反讽的分析和检测具有非常高的挑战性，仅从字面理解内容会得到完全相反的分析结果
- 立场分析：目标是识别出讨论或辩论双方的所持立场