摘要

英语语言中的情感分析研究近年来有了很大发展。然而，尽管中国电子商务和电子市场呈指数级增长，但中国的情绪分析研究并没有显著发展。本文旨在从单语和多语两个角度研究汉语情感分析的过去、现在和未来。首先介绍和总结了情感语料库和词汇库的构建。接着，通过三种不同的分类框架对汉语单语情感分类进行了调查。最后，介绍了基于多语言方法的情感分类。在对相关文献进行综述后，我们认为，对中文概念及其相互关系进行更人性化的(认知)表达，可以克服现有资源的不足，从而提高中文的研究水平。随着中文在网络上的不断扩展，中文情感分析正成为一个越来越重要的研究领域。尤其是概念层次的感觉分析，是一个激动人心但又富有挑战性的工作，这种研究领域的方向，对未来具有广阔的前景。

定义

情感分析的最终目标可以概括为识别给定文本的情感或观点标签。
根据最终标签的类型，问题通常分为情感分类和情感/主观性识别。尽管如此，这两个子问题在实现最终目标时有着相似的工作流程。

分类

汉语情感分析研究主要有两种方法: 单语方法和多语言方法。

- 单语方法
  侧重于执行典型的实时分析任务，如直接基于中文的极性检测。
- 多语言方法
  利用现有的英语资源和机器翻译技术来处理中文自然语言文本。

利用情感资源，研究路径分为基于机器学习和基于知识的方法。

- 机器学习将情感分类视为二进制(正或负)或多类分类问题。
- 基于知识的方法，研究语言规则和句法或语义关系。

机器学习方法

机器学习方法通常是一种有监督的方法，它不需要预定义的语义规则，而是需要一个有标签的数据集。它被重定向为一个文本分类问题。

- 第一步涉及提取特征
  这些特征通常分为以下几类:词汇特征、句法特征和语义特征。例如，否定标记、n-gram(单gram、双gram等。)、词性标签等等。
- 下一步是使用分类技术训练和测试数据
  如神经网络，最大熵，SVM和其他。

基于知识的方法

另一种流行的方法是基于知识的方法，或者通常称为无监督方法

混合方法

将机器学习方法与基于知识的方法相结合
例如：

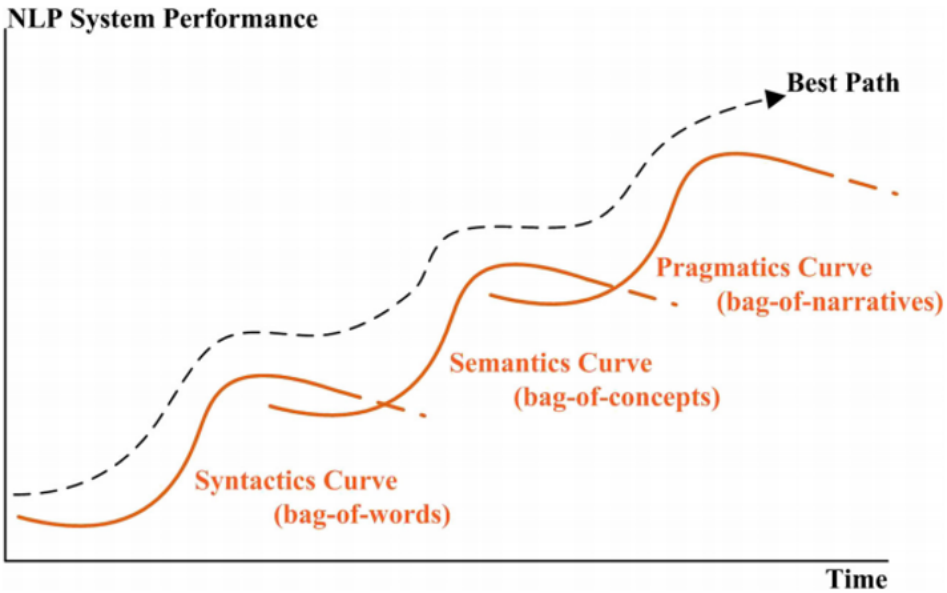
- 袁等36]用两种方法对微博情感进行分类。对于无监督方法，他们将简单情感词计数方法(SSWCM)与三个中文情感词典相结合。对于监督方法，他们测试了三个具有多种特征的模型(NB分类器、最大最小熵分类器和随机森林分类器)。他们的结果表明随机森林分类器在三个模型中提供了最好的性能。

多语言方法

万37]提出了一种方法，将英语和汉语分类结合起来。中文语言评论首先通过机器翻译被翻译成英文。然后对英文和中文评论进行分析，并结合它们的结果来提高情感分类的整体性能。上述方法的问题是，如果两种语言的领域知识不同，机器翻译的输出是不可靠的。这可能导致错误的积累，降低翻译的准确性。
因此，一些研究人员将这种情况归纳为一种领域适应性。

趋势

正如图6所示，自然语言处理研究正逐步从词汇语义学转向组合语义学。据我们所知，目前还没有关于概念层面的中文情感分析的工作。



语料库和词典的构建

- 语料库
- 汉语情感词汇表
  只包含感伤词:“Never- Ending Language Learner” (NELL)[46],
  同时包含感伤词和感伤极性:“National Taiwan University Sentiment Dictionary”(NTUSD)
  [47]“HowNet[48]”
  包含感伤词及其相关极性值:“SentiWordNet”[49]和“SenticNet[50]”

预处理

分词工具

有三种最流行的中文分词器可用:ICTCLAS、THULAC和Jieba分词器。

- ICTCLAS具有最高的精度，但速度最慢。

- Jieba是最快的分割器，但它的精度是最低的。
- 在速度和精度的权衡中，THULAC是三个分割器中最好的。

Table 1 Comparison between popular Chinese text segmentors

Algorithm	F-Measure		Speed	Supported Language
	msr_test (560KB)	pku_test (510KB)	CNKI_journal.txt (51MB)	
ICTCLAS(2015)	*0.891*	*0.941*	490.59KB/s	C, C++, C#, Java, Python
Jieba(C++)	0.811	0.816	*2314.89KB/s*	C++, Java, Python, R, etc.
THULAC_lite	0.888	0.926	1221.05KB/s	C++, Java, Python, SO

测试数据集

- 第一个数据集“ChnSentiCorp ”是由谭和张构建的。
  它包含三个领域的1021个文件:教育，电影和房子。同时，谭还收集了一个包含5000篇正面短文和5000篇负面短文的大规模酒店评论数据集。
- 第二个数据集，IT168TEST，是由扎吉巴洛夫和卡罗尔在[提出的产品评论数据集30]。
  该数据集包含超过20000条评论，其中78%被人工标记为阳性，22%被标记为阴性。
- 第三个数据集由第八届中国语言处理研讨会提供。
  这是一个基于主题的中文信息极性分类任务。
- 流行测试数据集的一些例子

Table 2 Some examples on popular testing datasets

	ChnSentiCorp		IT168TEST		SIGHAN8	
	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy
Li et al. [71]	–	–	–	–	–	63.51
Cao et al. [23]	–	–	–	–	71.00	–
Chen et al. [66]	–	63.43	–	–	–	–
Fu and Xu [78]	–	81.30	–	–	–	–
Zhang and He [35]	–	–	94.02	95.00	–	–
Zhai et al. [19]	88.60	88.60	80.90	81.30	–	–
Tan and Zhang [18]	88.58	-	–	–	–	–
Wan [37]	–	–	86.00	86.10	–	–
Wei and Pal [73]	–	–	–	85.40	–	–
Zagibalov and Carroll [30]	–	–	86.86	–	–	–

补充：https://chinesenlp.xyz/zh/docs/sentiment_analysis.html