第二章：资源能源相关互联网大数据分析技术综述 20

1. 预测分析技术
   1. 计量经济学方法

自回归模型（英语：Autoregressive model，简称AR模型），是统计上一种处理[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97" \t "https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B/_blank)的方法，被广泛运用在经济学、信息学、自然现象的预测上。

AR

其公式意为，X的当期值等于一个或数个落后期的线性组合，加常数项，加随机误差。自回归方法的优点是所需资料不多，可用自身变数数列来进行预测。在预测中应用最多的计量经济学模型是多元线性回归、基于经济学系统的联系方程、探讨多个经济变量动态关系的向量自回归（VAR）和向量误差修正模型（VECM）。

自回归只能适用于预测与自身前期相关的经济现象，即受自身历史因素影响较大的经济现象，如矿的开采量，各种自然资源产量等；对于受社会因素影响较大的经济现象，不宜采用自回归，而应改采可纳入其他变数的[向量自回归模型](https://baike.baidu.com/item/%E5%90%91%E9%87%8F%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B/_blank)。

向量[自回归模型](https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/_blank)（Vector autoregression,VAR)是AR模型的推广。VAR模型把系统中每一个[内生变量](https://baike.baidu.com/item/%E5%86%85%E7%94%9F%E5%8F%98%E9%87%8F" \t "https://baike.baidu.com/item/_blank)作为系统中所有内生变量的滞后值的函数来构造模型，从而将单变量自回归模型推广到由多元[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97" \t "https://baike.baidu.com/item/_blank)变量组成的“向量”自回归模型。VAR模型是处理多个相关经济指标与预测最容易操作的模型之一，并且在一定的条件下，多元MA和ARMA模型也可转化成VAR模型，因此近年来VAR模型受到越来越多的经济工作者的重视。

1.2 时间序列分析方法

时间序列分析方法中使用最多的是两类模型，一类是对序列自身变化情况进行刻画的自回归移动平均（ARMA）模型。这一类模型应用广泛，除了标准的ARMA之外，还有适用于非平稳序列ARIMA模型、以及考虑季节效应的季节自回归移动平均SARMA或者SARIMA模型。另一类是讨论序列波动情况的自回归条件异方差（ARCH）模型族。这类模型族中有很多种模型形式，如ARCH模型、GARCH模型、TARCH模型、EGARCH模型、非对称GARCH模型，等等。此外，还有一些预测模型融合了机制转移过程等，例如最常用的马尔科夫机制转换模型（Markov regime switch model）。

ARIMA模型全称为自[回归](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92/53132" \t "https://baike.baidu.com/item/_blank)积分滑动平均模型(Autoregressive Integrated Moving Average Model,简记ARIMA)，是由博克思(Box)和[詹金斯](https://baike.baidu.com/item/%E8%A9%B9%E9%87%91%E6%96%AF" \t "https://baike.baidu.com/item/_blank)(Jenkins)于70年代初提出一著名时间序列预测方法 ，所以又称为box-jenkins模型、博克思-詹金斯法。ARIMA模型的基本思想是：将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列。这个模型一旦被识别后就可以从时间序列的过去值及现在值来预测未来值。现代统计方法、[计量经济模型](https://baike.baidu.com/item/%E8%AE%A1%E9%87%8F%E7%BB%8F%E6%B5%8E%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/_blank)在某种程度上已经能够帮助企业对未来进行预测。

1.3 人工智能和机器学习的数据挖掘方法

逻辑回归

逻辑回归（logistic回归）是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此[因变量](https://baike.baidu.com/item/%E5%9B%A0%E5%8F%98%E9%87%8F" \t "https://baike.baidu.com/item/_blank)就为是否胃癌，值为“是”或“否”，自变量如年龄、性别、饮食习惯、[幽门螺杆菌](https://baike.baidu.com/item/%E5%B9%BD%E9%97%A8%E8%9E%BA%E6%9D%86%E8%8F%8C" \t "https://baike.baidu.com/item/_blank)感染等。自变量既可以是连续的，也可以是分类的。然后通过logistic回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

逻辑回归使用 Sigmoid 函数，其形式为：



支持向量机

支持向量机是由Vapnik领导的AT&T Bell实验室研究小组在1995年提出的一种新的非常有潜力的分类技术，它在解决小样本、非线性及[高维](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4" \t "https://baike.baidu.com/item/_blank)模式识别中表现出许多特有的优势，并能够推广应用到[函数](https://baike.baidu.com/item/%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)[拟合](https://baike.baidu.com/item/%E6%8B%9F%E5%90%88" \t "https://baike.baidu.com/item/_blank)等其他[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0" \t "https://baike.baidu.com/item/_blank)问题中。

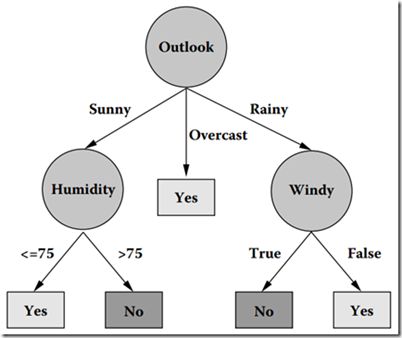
支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的[超平面](https://baike.baidu.com/item/%E8%B6%85%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)的两边建有两个[互相平行](https://baike.baidu.com/item/%E4%BA%92%E7%9B%B8%E5%B9%B3%E8%A1%8C" \t "https://baike.baidu.com/item/_blank)的超平面。建立方向合适的分隔[超平面](https://baike.baidu.com/item/%E8%B6%85%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)使两个与之平行的超平面间的距离最大化。其假定为，平行[超平面](https://baike.baidu.com/item/%E8%B6%85%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)间的距离或差距越大，分类器的总误差越小。

SVM的关键在于核函数。低[维空间](https://baike.baidu.com/item/%E7%BB%B4%E7%A9%BA%E9%97%B4" \t "https://baike.baidu.com/item/_blank)向量集通常难于划分，解决的方法是将它们映射到[高维空间](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4%E7%A9%BA%E9%97%B4" \t "https://baike.baidu.com/item/_blank)。但这个办法带来的困难就是计算[复杂度](https://baike.baidu.com/item/%E5%A4%8D%E6%9D%82%E5%BA%A6" \t "https://baike.baidu.com/item/_blank)的增加，而[核函数](https://baike.baidu.com/item/%E6%A0%B8%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)正好巧妙地解决了这个问题。也就是说，只要选用适当的[核函数](https://baike.baidu.com/item/%E6%A0%B8%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)，就可以得到[高维空间](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4%E7%A9%BA%E9%97%B4" \t "https://baike.baidu.com/item/_blank)的分类函数。在SVM理论中，采用不同的[核函数](https://baike.baidu.com/item/%E6%A0%B8%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)将导致不同的SVM算法。

决策树与随机森林

决策树是一种基本的分类和回归方法，分类决策树如 ID3、C4.5 算法[35,36] 等，回归决策树如 CART 算法[37]等。在回归分析杠杆影响因素中，需要使用回 归决策树。CART 本身即可用于分类也可用于回归，它假设决策树是二叉树，对 回归树用平方误差最小化准则，对分类树用基尼指数（Gini index）最小化准则， 由生成和剪枝两步组成。在回归树生成过程中，采用启发式的方法选择变量和它 的取值，作为切分变量和切分点，并寻找最优的切分变量和切分点递归地生成最 小二乘树。 决策树是随机森林的基础，用于分类的决策树中，ID3 算法核心是在决策树 各个结点上使用信息增益选择特征，C4.5 算法对 ID3 算法进行了改进，使用信 息增益比来选择特征。

IMG_256



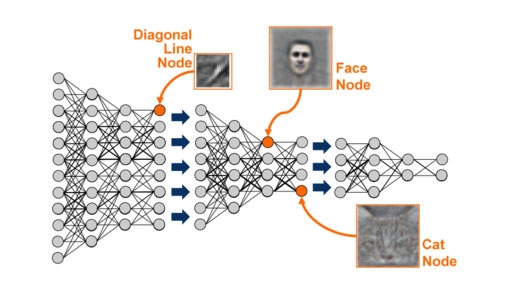
鉴于决策树容易产生过拟合，随机森林采用多个决策树投票机制进行改善 [39]。RF 生成首先从样本集中采用 Bootstrap 方式产生 N 个样本，然后选取 A 个特征中的 k 个构建决策树，重复 M 次产生 M 棵决策树后，采用多数投票机制进 行预测。 随机森林的主要思想是每一棵决策树是精通于某一个窄领域的专家（因为从 A 个特征中选择 k 让每一棵决策树进行学习），这样在随机森林中就有了很多个 精通不同领域的专家，对一个新的问题（新的输入数据），可以用不同的角度去 看待它，最终由各个专家，投票得到结果。 随机森林的优点在于能够处理很高维度的数据并且不用做特征选择，在训练 完成后它也能给出不同特征的重要性，并且容易理解，具有很好的抗噪声能力， 可以进行并行处理。

深度学习

[深度学习](https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)的概念源于[人工神经网络](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)的研究。含多隐层的[多层感知器](https://baike.baidu.com/item/%E5%A4%9A%E5%B1%82%E6%84%9F%E7%9F%A5%E5%99%A8" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。

深度学习的概念由Hinton等人于2006年提出。基于深度置信网络(DBN)提出非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来希望，随后提出多层自动编码器深层结构。

[深度学习](https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)是机器学习中一种基于对数据进行[表征](https://baike.baidu.com/item/%E8%A1%A8%E5%BE%81" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)学习的方法。观测值（例如一幅图像）可以使用多种方式来表示，如每个[像素](https://baike.baidu.com/item/%E5%83%8F%E7%B4%A0" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)强度值的向量，或者更抽象地表示成一系列边、特定形状的区域等。而使用某些特定的表示方法更容易从实例中学习任务（例如，[人脸识别](https://baike.baidu.com/item/%E4%BA%BA%E8%84%B8%E8%AF%86%E5%88%AB" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)或面部表情识别）。深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取[特征](https://baike.baidu.com/item/%E7%89%B9%E5%BE%81" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)。

IMG_256

深度学习在分类问题如图像、NLP 领域 取得了越来越多的成果，其激活函数通常采用 Sigmoid 函数、tanh 函数、ReLU 函数等，输出层常用 Sigmoid 或 Softmax 层，训练算法主要采用梯度下降法如 BP 算法[40]。常见的神经网络模型有深度置信网络（Deep Belief Network, DBN）、 深度玻尔兹曼机（Deep Boltzmann Machine, DBM）、含有卷积层和池化层的卷积神经网络（Convolutional Neural Network, CNN）、循环神经网络（Recurrent Neural Networks, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM） 等。深度学习应用领域跨度较大，如图像识别主流算法 CNN，语音识别和机器 翻译，文本模型、时序数据处理以及推荐系统中都有广泛的应用。对极端亏损情况的分类尝试使用合适的网络结构进行分析预测。

深度学习源于人工神经网络的研究。人工神经 网络(artificial neural network, ANN)作为计算工具是 由文献[1]引入。之后，Hebb自组织学习规则、感知 机模型、Hopfield神经网络、玻尔兹曼机、误差反向 传播算法和径向基神经网络等也相继被提出。文献 914 电 子 科 技 大 学 学 报 第 46 卷 [2]利用逐层贪心算法初始化深度信念网络，开启了 深度学习的浪潮，指出深度学习的本质是一种通用 的特征学习方法，其核心思想在于提取低层特征， 组合形成更高层的抽象表示，以发现数据的分布规 律。文献[2]的方法有效地缓解了DNN层数增加所带 来的梯度消失或者梯度爆炸问题。随后文献[3]使用 自动编码机取代深度信念网络的隐藏层，并通过实 验证明了DNN的有效性。同时，研究发现人类信息 处理机制需要从丰富的感官输入中提取复杂结构并 重新构建内部表示，使得人类语言系统和感知系统 都具有明显的层结构[4]，这从仿生学的角度，为DNN 多层网络结构的有效性提供了理论依据。 此外，深度学习的兴起还有赖于大数据和机器 计算性能的提升。大数据是具有大量性、多样性、 低价值密度性的数据的统称，深度学习是处理大数 据常用的方法论，两者有紧密的联系。以声学建模 为例，其通常面临的是十亿到千亿级别的训练样本， 实验发现训练后模型处于欠拟合状态，因此大数据 需要深度学习[5]。另外，随着图形处理器(graphics processing unit, GPU)的发展，有效且可扩展的分布 式GPU集群的使用大大加速了深度模型的训练过 程，极大地促进了深度学习在业界的使用。

1. 文本处理相关技术 1页半

自然语言处理技术综述 500字

自然语言处理是[计算机](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA" \t "https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/_blank)科学领域与[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD" \t "https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/_blank)领域中的一个重要方向。它研究能实现人与[计算机](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA" \t "https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/_blank)之间用自然语言进行有效通信的各种理论和方法，综合了语言学、计算机科学、逻辑学、心理学、 人工智能等学科的知识与成果。其主要研究任务包括词性标注、机器翻译、命名实体识别、机器问答、情感分析、自动文摘、句法分析和共指消解等。

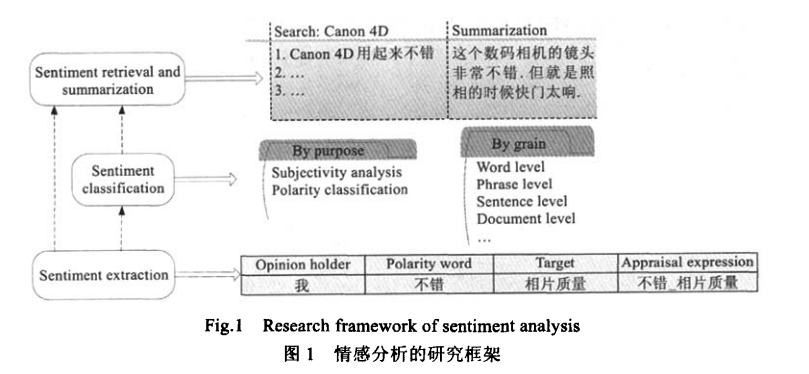
自然语言作为高度抽象的符号化系统，文本间的关系难以度量，相关研究高度依赖人工构建特征。而深度学习方法的优势恰恰在于其强大的判别能力和特 征自学习能力，非常适合自然语言高维数、无标签和大数据的特点。当前，如何应用深度学习技术解决自然语言处理(NLP)相关任务是深度学习的研究热点。2013年，随着词向量word2vec[6]的兴起， 各种词的分布式特征相关研究层出不穷。2014年开 始，研究者使用不同的DNN模型，例如卷积网络， 循环网络和递归网络，在包括词性标注、情感分析、 句法分析等传统NLP应用上取得重大进展。2015年 后，深度学习方法开始在机器翻译、机器问答、自动文摘、阅读理解等自然语言理解领域攻城略地， 逐渐成为NLP的主流工具。目前，深度学习用于NLP领域的主要步骤可以 归结为如下3步：1) 将原始文本作为输入，自学习得到文本特征的分布表示。2) 将分布式向量特征作 为深度神经网络的输入。3) 针对不同的应用需求， 使用不同的深度学习模型，有监督的训练网络权重。在未来几年，深度学习将持续在自然语言理解领域做出巨大影响。

文本情感分析技术

文本情感分析又称意见挖掘,简单而言,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的 过程.最初的情感分析源自前人对带有情感色彩的词语的分析,如,“美好”是带有褒义色彩的词语,而“丑陋”是 带有贬义色彩的词语.随着互联网上大量的带有情感色彩的主观性文本的出现,研究者们逐渐从简单的情感词 语的分析研究过渡到更为复杂的情感句研究以及情感篇章的研究。

按照处理文本的类别不同,可分为基于新闻评论的情感分析和基于产品评论的情感分析.其中,前者处理的文本主要是新闻评论,如情感句“他坚定 地认为台湾是中国不可分割的一部分”,表明了观点持有者“他”对于事件“台湾归属问题”的立场;后者处理的主 要是网络在线的产品评论文本,如“Polo 的外观很时尚”,表明了对评价对象“Polo 的外观”的评价“时尚”是褒义 的.基于产品评论的情感分析可以帮助用户了解某一产品在大众心目中的口碑,对于判断消费者喜好和商家营业情况、行业景气程度等有重要作用.而基于新闻评论的情感分析多用于舆情监控和信息预测中.

情感分析可归纳为 3 项层层递进 的研究任务,即情感信息的抽取、情感信息的分类以及情感信息的检索与归纳。



1.情感信息抽取

情感信息抽取是情感分析的最底层的任务,它旨在抽取情感评论文本中有意义的信息单元.其目的在于将无结构化的情感文本转化为计算机容易识别和处理的结构化文本,继而供情感分析上层的研究和应用服务.如将情感句“我觉得 Canon 的相片质量不错”转化为如图 1 所示的结构化文本形式.情感信息抽取的内容包括评价词语的抽取和判别、评价对象的抽取、观点持有者抽取以及组合评价单元的抽取，需要使用的技术包括属性抽取、规则/模板法、句法分析、序列标注等。

评价词语又称极性词、情感词,特指带有情感倾向性的词语。词语的情感极性判别主要有基于语料库和基于词典两种方法。基于语料库的方法主要是利用词语之间的连词以及统计特征来判别词语的情感极性。由连词连接的词语的情感极性 存在某种关联，比如由连词“和”连接的词语的情感极性相 同，由连词“但是”连接的词语的情感极性相反。基于词典的方法是利用WordNet 或 HowNet提供的语义相 似度或者层次结构来判别词语的情感极性。

2.情感信息分类

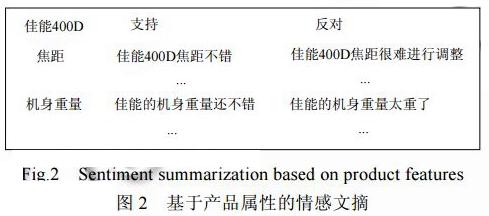
情感信息分类则利用底层情感 信息抽取的结果将情感文本单元分为若干类别,供用户查看,如分为褒、贬两类或者其他更细致的情感类别(如 喜、怒、哀、乐等).按照不同的分类目的,可分为主客观分析和褒贬分析;按照不同的分类粒度,可分为词语级、 短语级、篇章级等多种情感分类任务。

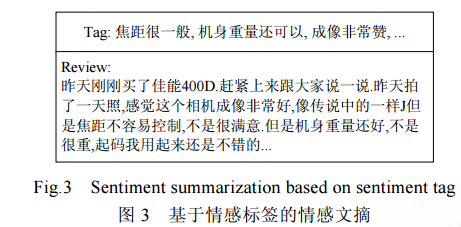
其中最常见的是主观信息的情感分类,包括最常见的褒贬二元分类以及更细致的多元分类，主要方法为基于情感知识的方法以及基于特征分类的方法。前者主要是依靠一些已有的情感词典或领域词典以及主观文本中带有情感极性的组合评价单元进行计算,来获取 主观文本的极性。后者主要是使用机器学习的方法,选取大量有意义的特征来完成分类任务。代表性的研究工作首先分析句子/篇章中的评价词语或组合评价单元的极性,然后进行 极性加权求和.这种方法的重点一般都放在评价词语或组合评价单元的抽取和极性判断方法的研究上。次将机器学习的方法应用于篇章级的情感分类任务，研究重点在于有效特征的发现以及特征选择和特征融合等问题的研究。

如情绪评估工具Opinion Finder 和 GPOMS。Opinion Finder 根据文本内容可以评估“积极”与“消极”两种情绪；GPOMS 根据文本内容可以评估 “calm(冷静)”、“alert(警觉)”、“sure(确信)”、“vital(活泼)”、“kind(美好)”、“happy(高 兴)”六种情绪

3.情感信息的检索与归纳

最后，情感分析技术与用户的交互主要集中于情感信息检索和情感信息归纳两项任务上。情感信息检索要求检 索回的文档同时满足两项准则:(1) 主题相关;(2) 具有情感倾向性。传统的基于事实性新闻语料的文摘旨.相比而言,情感文摘的处理对象为某一产品或某一事件的大量用户评论,在提取重要的事实性信息,并去除冗余信息的基础上融入了更多的情感信息。它主要侧重于提取具有明显情感倾向性的主观信息,是对某一产 品或某一事件的评论信息的归纳和汇总。针对产品类评论信息,情感文摘共有两种呈现方式:一种是基于产品属性的情感文摘,另一种是基于情感标签的情感文摘。





1.4 集成学习技术