第 3 章

电子商务产品质量评论观点识别及情感倾向分析

随着电子商务的快速发展，越来越多的人通过电子商务网站来了解产品信息、购买商品，并且通过评价表达自己购买商品过程中的感受、对购买商品的满意程度和相关建议要求。评价和打分等商品舆情信息是买家了解电子商务网站产品和商家服务的一种重要渠道和表达方式[60]。文本情感分析是对给出的文本的感情色彩进行分析、归纳的过程[61]，即判断一篇文本中观点持有者对某个事件或商品持有的正向、负向或中立的态度。它属于信息检索或者自然语言处理的范畴。目前，国内已有许多专家、学者根据实现的方法将舆情分析技术分为基于词的倾向性分析和基于机器学习的倾向性分析[62]。例如，杨震等人在网络舆情内容分析中，提出基于字符串相似性聚类的网络短文本舆情热点发现技术[63]。Kouloumpis 等利用微博中的口语和网络语言来提高情感倾向分析的准确性[64]。

3.1 电子商务产品质量情感倾向词典构建

在林鸿飞教授等[65]所构建的中文情感词汇本体库的基础上构建出对电子商务产品质量主题针对性强的情感倾向词典，包括电子商务产品质量主题领域词汇和对应的网络用语词汇，最终构建情感倾向词典，能更加全面地对电子商务产品质量数据进行情感倾向分析。

3.1.1 中文分词方法

中文不像英文那样每个词汇之间由空格分开，需先进行分词才能进一步处理。本研究采用最大匹配算法对中文文本进行分词，该方法属于基于字符串匹配的分词方法，需要分词词典支持。分词词典本研究采用中科院 ICTCLAS 分词系统[66]，该词典搜集了日常生活中使用频率较高的 56 008 个词汇，基本能够满足分词的需要。

在特征选择方法上，本研究采用了情感词典作为特征选择的依据，所以在分词过程中，将与电子商务产品质量主题领域相关的词汇以及网络用语添加到分词系统 ICTCLAS 的词典中，其中最大匹配的步长设置为四个汉字，只对中文内容进行分词处理，将其并集作为分词词典的结果，更加有效合理。

本研究提出使用情感词典对文本进行表示，这个过程在中文分词阶段就能完成，不需要单独的特征选择步骤。文本处理流程如图 3.1 所示。

在构建适合于电子商务产品质量话题型的领域情感词典

时，需要从评论中获取和话题相关的领域性词汇。为此，本研究通过搜狗细胞词库获取和话题相关的领域词汇，在对电子商务产品质量话题型评论文本进行分词中，向中科院分词系统加入领域词汇，进一步词频分析，然后通过预处理删除一些无关的字词和符号，最后通过和已有的情感词典进行匹配，筛选出和话题相关的领域词汇。

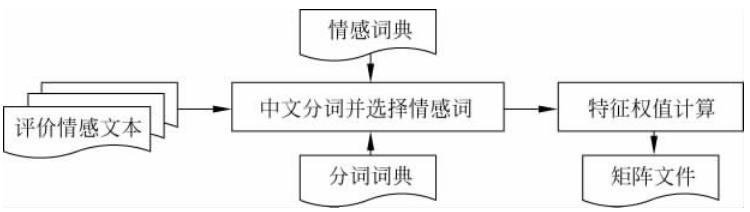

图 3.1　文本处理流程

3.1.2　领域情感词典构建

由于在不同话题微博评论中，往往会出现很多和微博话题相关的情感词汇，这些词汇不包含在基础情感词典中，但是却富含和情感相关的信息，对微博评论情感分析具有很重要的影响。例如，“#房价问题#现在房价太高了，有的新房没多少年就坍塌了！”，这句关于房价问题的微博话题评论中的“坍塌”词汇是基础情感词典中不具备的，但是却明显表明了关于房子的态度和想法。因此，本研究从搜狗细胞词库以及互联网搜集常见富含情感的网络用语，构建适合电子商务产品质量话题的领域情感词典。

在构建领域词典时，同样需要利用基础情感词典计算领域词语的情感倾向和情感强度，具备更多情感词数量的基础情感词典将有效地提高领域词典的构建精确度。因此，本研究在林

鸿飞教授等人所构建中文情感词汇本体库的基础上，对其进行修改和调整，构建适合话题型电子商务产品质量的基础情感词典。在情感词汇本体中，一般的格式如表 3.1 所示。

表 3.1 情感词汇本体格式举例

词语	词性种类	词义数	词义序号	主要情感分类	主要强度	主要极性	辅助情感分类	辅助强度	辅助极性
脏乱	Adj	1	1	NN	7	2	—	—	—
臭名昭彰	idiom	1	1	NN	9	2	—	—	—
周到	adj	1	1	PH	5	1	—	—	—
言过其实	idiom	1	1	NN	5	2	—	—	—

在构建情感倾向词典时，本研究采用中文情感词汇本体库的情感分类、情感强度两个维度，将基础情感词情感极性分为三类：正面情感、中立情感、负面情感。在情感强度中，中立情感用 0 表示，正面情感用正号＋表示，负面情感用负号－表示，情感强度分为－9，－7，－3，－1，0，1，3，5，7，9，其中 9 表示正面情感倾向程度最大，－9 为负面情感倾向程度最大。由于中文情感词汇本体库缺乏中性情感词，本研究收集相关中性词语加入中文情感词汇本体库。最终得到正面情感词语 10 541 个，负面情感词语 10 102 个，中性情感词语 4127 个。具体示例如表 3.2 所示。

表 3.2 基础情感词典举例

极性	权值	基础情感词示例
正面	[1，9]	雅兴、怡悦、致敬、敬佩、高兴、喜欢
负面	[－9，－1]	脏乱、糟糕、早衰、责备、悲伤、哭泣
中立	0	一般、中立、平庸、无功无过、平淡

由于在汉语中，很多词语（多数为动词和形容词）存在一词

多义的现象，在不同的话题领域中，一些词语的语义和情感极性有所不同。需要采取相应的方法，减少这方面因素影响情感分类精确度。例如，卢苇提出构建不受领域主题影响的中文基础情感词典[67]。但是这样的方法有一定缺陷：构建不受主题领域影响的基础情感词典，将会导致大量情感词被排除在外，导致基础情感词典过小，需要大量的人工参与。通过对话题型评论文本的研究，发现针对某一话题评论时，受话题领域影响的基础情感词是很小部分的。因此，为了减少人工参与量，并且增加基础情感词典中的基础词语数量，本研究针对话题型评论情感分析研究时，根据不同的话题，通过词频分析结合人工识别找出受该话题影响的词语，对基础情感词典进行一些调整。

3.1.3 程度副词词典构建

程度副词是副词组成之一，主要用于修饰动词和形容词，改变词语情感的强弱。大多数用户直接用情感词表达观点和情感，并且常常使用程度副词来加强或减弱自己的情感。因此，程度副词也是影响情感的重要情感特征项之一。例如评论："#房价问题#房价有一点点高"，评论中，"一点点"程度副词影响了评论句中的观点和情感。由此可见，程度副词的使用确实影响了评论中的情感倾向程度。本研究在构建程度副词词典时，参考游建平等人对程度副词的四个分类：低量、中量、高量、极量，选用知网提供的中文程度级别词语，一共 219 个[68]。同时参考宋静静对程度副词的权值设置进行改进，将程度副词权值范围设置为[0.5,2]，最终得到低量级别 41 个、中量级别 37 个、高量级别 42 个、极量级别 99 个[69]。具体示例如表 3.3 所示。

表 3.3 程度副词举例

级别	权值	程度副词示例
低量	0.5	多多少少、略加、一点、有些、稍许
中量	1.0	进一步、较为、更加、愈发、越
高量	1.5	多多、分外、实在、特别、尤其
极量	2	过分、过猛、极度、非常、绝对

3.1.4 否定词词典构建

否定词是对行为或状态进行否定的副词。主要用于修饰动词、形容词。文本中出现否定词，将会影响被修饰情感词的极性。例如，若否定词个数为 $2a+1$ 个，则被修饰情感词的极性将会相反；若否定个数为 $2a$，则被修饰情感词的极性不变。在电子商务产品评论中，网民经常使用否定词来支持或否定一些事物。例如，“#房价问题#房价又涨了，很不高兴!”这句话中“高兴”表达正面情感，但是用否定词“不”修饰“高兴”后，这条关于房价问题的评论情感从正面情感转变为负面情感。因此，在分析电子商务产品评论情感时，需要构建合理的否定词词典，并赋予其权值为－1。本研究采用郝雷红提出的 31 个否定副词[70]。具体示例如表 3.4 所示。

表 3.4 否定副词举例

否定副词示例	权值	个数
白、甭、别、不、不必、不曾、不要、不用、非、干、何必、何曾、何尝、何须、空、没、没有、莫、徒、徒然、枉、未、未曾、未尝、无须(无须乎、无需、毋须)、毋庸(无庸)	－1	31

3.1.5 网络用语词典构建

网络用语伴随着网络的发展而兴起，大量的网络词汇诞生，被广大网民熟知和使用。电子商务平台作为一种新兴社交媒体，已成为网民传播信息最为火热的工具。由于网络语言的魅力，电子商务产品质量评论文本包含大量网络用语，而这些网络用语往往具有强烈的情感倾向。尤其在话题型微博评论中，绝大多数网民更加倾向使用具有情感性的网络词汇。目前网络用词的类型有数字型、字母型、同音型等。例如，正面情感的网络用语有狂顶、大神、hold 住等，负面情感的网络用语有菜鸟、555等。本研究从搜狗细胞词库以及互联网搜集常见富含情感的网络用语，最终采用人工判断的方法，给网络权值赋值，设置权值范围[－9，9]。具体示例如表 3.5 所示。

表 3.5 网络用语举例

极性	网络用语示例	权值	数目
正面情感	大神、hold 住、完爆、我顶	[1，9]	85
负面情感	菜鸟、呜呜、弱爆了	[－9，－1]	106

3.2 电子商务产品质量话题评论情感倾向分析

本研究采用情感词典对电子商务产品质量评论文本进行情感分类，建立高质量的情感词典，有效地保留了情感相关特征项之间的关系，考虑了情感词本身存在情感强度的因素。

3.2.1 文本情感特征项抽取算法

本研究在上下文滑动算法基础上，将词性规则、情感词典、平滑算法相结合，对电子商务产品质量话题型评论的情感相关特征项（情感词、程度副词、否定词、表情符号、网络用语、评价对象）进行抽取。

经过分词之后，评论文本转变为词汇序列串。上下文滑动窗口，是指在上下文环境中，以某一词为中心，向前和向后推进 n 个字或词，形成一个队列缓存区。通过上下文滑动窗口，考查词的词法层特征，包括局部词、局部词性、局部共现、词类搭配等。若文本为 $\{t_1, t_2, \cdots, t_m\}$ $(m \geqslant 2n+1)$，t 表示文本中的词语，则以词语 W 为中心窗口建立大小为 n 的上下文滑动窗口，左窗口 LW 可表示为 $(\mathrm{LW}_1, \mathrm{LW}_2, \cdots, \mathrm{LW}_n)$，右窗口 RW 可表示为 $(\mathrm{RW}_1, \mathrm{RW}_2, \cdots, \mathrm{RW}_n)$。由于情感词的修饰词不会超过三个，故文本设定滑动窗口大小为 3。

在具体抽取与情感词相关的否定词、程度副词时，面对评论中时常出现多个情感词的情况，文本以词性规则锁定某一情感词，然后通过上下滑动算法，对其相关的否定词和程度副词进行抽取，以适应具体特征性抽取环境。

1. 情感词、程度副词、否定词和评价对象的抽取

在文本中程度副词和否定词为情感词的修饰词语，由于这两类修饰词通常离情感词最近，对情感词有重要影响，因此，文本采用上下文滑动窗口来抽取评论中每一个情感词组合单元时，设定滑动窗口大小的取值为 3。具体算法如表 3.6 所示。

表 3.6 情感词、程度副词、否定词和评价对象的抽取算法

输入：话题评论文本集合 $D=\{D_1,D_2,D_3,\cdots,D_n\}$，词典资源（情感词典、程度副词词典、否定词典、评价对象词典），四个标点集合{，，。，！，？}
输出：情感词组合单元集合 T 及其评价对象集合 E_O
特征抽取算法描述：
(1) 循环取出评论 $D_i \in D$
(2) 将 D_i 根据标点集合划分为 j 份评论
(3) 根据一般用语习惯，在查找情感词时，从左到右查找情感词。如果 D_{ij} 评论包含词性为动词或者名词的词语，则通过情感词典进行匹配，假设找到 m 个情感词，标记位置，获取其权值，并记住每个情感词位置，将第 k 个情感词标记为 EW_{ijk} 中心
(4) 以情感词 EW_{ijk} 为中心抽取程度副词、评价对象过程中，设置以下规则：如果抽取到程度副词、评价对象，那么抽取相应的特征项结束，接下来抽取其他特征项，或者利用平滑算法抽取，遇到其他情感词则换个方向抽取特征项
(5) 以情感词 EW_{ijk} 为中心，在 D_{ijk} 内抽取程度副词、否定词、评价对象。采用上下文滑动算法，左右距离为 3，按照一般用语习惯，从情感词的左到右，使用程度副词词典、否定词典抽取副词 AW_{ijk}、否定词 PN_{ijk}，并获取程度副词的权值以及否定词的个数。将 D_{ij} 的情感组合单元添加到集合 T_{ij} 中
(6) 对 D_{ij} 抽取评价对象时，结合使用词性规则，设置以下规则：如果情感词 EW_{ijk} 的词性为动词，则采用上下文滑动算法，从情感词的右到左查找，利用评价对象词典，对词性为名词的评价对象 O_{ijk} 进行抽取，并获取对应的权值，添加到集合 E_O 中；如果情感词 EW_{ijk} 的词性为形容词，则采用上下文滑动算法，从情感词的右到左查找，利用评价对象词典，对词性为名词的评价对象进行抽取，并获取对应权值。在对 D_{ij} 评论内容进行抽取对象时，会遇到两个评价对象或者无评价对象的情况，设置评价对象的选择规则：①如果以词性为形容词的情感词 EW_{ijk} 为中心，抽取到两个权值极性相反的评价对象时，选择第一个抽取到的评价对象，并将该评价对象 O_{ij} 添加到集合 E_O。②如果以词性为动词的情感词 EW_{ijk} 为中心，抽取到两个权值极性相反的评价对象时，选择第二个抽取到的评价对象，并将该评价对象 O_{ij} 添加到集合 E_O。③如果以情感词 EW_{ij} 为中心，抽取到两个权值极性一样的评价对象时，选择权值较小的评价对象。④如果在第 D_{ij} 份评论内容中没有指明评价对象，那么默认其评价对象为第 $i-1$ 份评论内容中的评价对象。如果 i 为 1，且没有评价对象，那么默认评价对象为话题本身

2. 网络用语的抽取

网络用语经过分词后，通过网络用语词典，对分词后的每条评论进行匹配抽取网络用语并得到权值。具体算法如表 3.7

所示。

表 3.7 网络用语的抽取算法

输入：话题评论文本集合 $D=\{D_1,D_2,D_3,\cdots,D_n\}$，词典资源（网络用语词典）
输出：网络用语集合 NL
(1) 循环取出评论 $D_1\in D$
(2) 使用网络用语词典匹配 D_1，抽取网络用语 NL_i 并获取其权值，添加到网络用语集合 NL

3. 表情符号的抽取

表情符号经过分词后，通过表情符号词典，对分词后的每条评论进行匹配抽取表情符号并得到权值。具体算法如表 3.8 所示

表 3.8 表情符号的抽取算法

输入：话题评论文本集合 $D=\{D_1,D_2,D_3,\cdots,D_n\}$，词典资源（表情符号词典）
输出：表情符号集合 EM
(1) 循环取出评论 $D_1\in D$
(2) 使用表情符号词典匹配 D_1，抽取表情符号 EM_i 并获取其权值，添加到表情符号集合 EM

电子商务产品质量话题型评论经过预处理、分词以及词性标注等情感特征项抽取处理后，通过平滑算法、结合词性规则、情感词典三者相结合，以情感词为中心，逐渐提取情感词、否定词、程度副词、评价对象、网络用语等情感特征项。

3.2.2 语句情感特征权值计算

特征权值是指特征词在文本中的权重，也可称为词的向量，

是分类器分类的重要依据。本研究使用词频、布尔型(Boolean)两种权值进行情感分类对比。一般在分词处理完成后就可以计算特征权值,然后特征选择后输入分类器。使用情感词典作为特征选择时,因为分词时可以完成特征选择,所以特征权值计算在特征选择之后进行。

该模块的主要功能是计算电子商务产品质量语句的情感倾向值。在情感计算过程中,每条电子商务产品质量语句情感由情感词的情感和表情符号的情感构成。在情感词情感计算中,否定词对情感词存在正反意义的作用,若否定词个数为 $2a+1$ 个,则用相反意义的词汇替代;若否定个数为 $2a$,则情感词不变,程度副词对情感词的情感强弱具有增减作用。由于表情符号和情感词一样能体现情感倾向,因此将文字句子的权重 α 取为 0.5,表情符号的权重 β 取为 0.5。本研究通过改进陈晓东[71]提出的微博情感倾向计算公式,得到每条电子商务产品质量评论的情感值计算公式如下所示:

$$S=\alpha\left(\sum_{1}^{m}(-1)^{N_j}C_jM_j+\sum_{1}^{w}W_k\right)+\beta\left(\sum_{1}^{s}O_i\right) \quad (3.1)$$

其中,m 为情感词个数,M_j 为该条电子商务产品质量评论中第 j 个情感词,C_j 为修饰情感词 M_j 的程度副词,N_j 为修饰情感词 M_j 的否定词,w 为网络用语个数,s 为表情符号个数。本研究将每条电子商务产品质量语句情感值计算结果分为三大类:正面情感倾向、中立倾向、负面情感倾向。

3.2.3 电子商务产品质量评论情感倾向计算

本研究选取和电子商务产品质量情感有关的特征项,获取每个特征项相应的权值,最后作求和运算,得到每条评论的情感

倾向值，从而判断其情感倾向。在情感计算过程中，将每条电子商务产品质量语句情感分为两部分构成：一部分为文字表述情感；另一部分为表情符号情感。文字表述情感包括情感词及其修饰词构成的情感和网络用语的情感。表情符号的情感包括表情图片和输入法表情符号。对评论文本进行数据预处理之后，本研究抽取了情感词、否定词、程度副词、评价对象、网络用语、表情符号情感特征项，并通过构建好的词典获取特征项的权值。

本研究对每条话题电子商务产品质量评论，按照标点符号进行分割，假设分割成 n 个句子，即一条评论 D_i 将会有 n 个句子 $D_{i1}, D_{i2}, D_{i3}, \cdots, D_{in}$，那么评论 d_i 的情感值由 n 个句子的情感值构成，并且在 D_{in} 内只选择一个评价对象。

在情感词情感计算中，否定词对情感词存在正反意义的作用，若否定词个数为 $2a+1$ 个，则用相反意义的词汇替代；若否定个数为 $2a$，则情感词不变，程度副词对情感词的情感强弱具有增减作用，评价对象对评论的极性也存在影响。在 D_{ij} 句子中，$1 \leqslant j \leqslant n$，情感词个数为 m，网络用语为 NL_i，表情符号为 EM_i，评价对象为 O_{ij}，情感词为 EW_{ij}，程度副词为 AW_{ij}，否定词个数为 PN_{ij}。D_{ij} 句子中第 $k(1 \leqslant k \leqslant n)$ 个情感词情感值的计算公式如下：

$$\mathrm{WE}_{ijk} = [(-1)^{\mathrm{PN}_{ijk}} \mathrm{AW}_{ijk} \times \mathrm{EW}_{ijk}] \tag{3.2}$$

其中，EW_{ijk} 表示第 k 个情感词的权值，AW_{ijk} 表示修饰第 k 个情感词的程度副词的权值，PN_{ijk} 表示修饰 k 个情感词的否定词的个数。

D_{ij} 句子中文字表述情感值的计算公式如下：

$$\mathrm{WE}_{ij} = O_{ij} \sum_{k}^{m} [(-1)^{\mathrm{PN}_{ijk}} \mathrm{AW}_{ijk} \times \mathrm{EW}_{ijk}] \tag{3.3}$$

其中，O_{ij} 表示 D_{ij} 句子的评价对象，m 表示情感词的个数。

D_i 句子文字表示情感值的计算公式如下：

$$\mathrm{WE}_i = \sum_{j}^{n} O_{ij} \left(\sum_{k}^{m} (-1)^{\mathrm{PN}_{ijk}} \mathrm{AW}_{ijk} \times \mathrm{EW}_{ijk} \right) + \sum_{s}^{w} \mathrm{NL}_{is} \tag{3.4}$$

其中，n 表示 D_i 分为 n 个字句，m 表示 D_i 句子中表情符号的个数，NL_{is} 表示第 s 个网络用语。

本研究认为表情符号和情感词一样能体现情感倾向，因此，将句子中文字表示情感值的权重 α 取为 0.5，表情符号的权重 β 取为 0.5。每条话题电子商务产品质量评论 D_i 句子的总情感值计算公式如下：

$$\mathrm{WE}_i = \alpha \left\{ \sum_{j}^{n} O_{ij} \left[\sum_{k}^{m} (-1)^{\mathrm{PN}_{ijk}} \mathrm{AW}_{ijk} \times \mathrm{EW}_{ijk} \right] + \sum_{s}^{w} \mathrm{NL}_{is} \right\} + \beta \sum_{h}^{z} \mathrm{EM}_{ih} \tag{3.5}$$

3.3 电子商务产品质量话题观点识别

目前观点句识别的方法主要还是采用机器学习。基于机器学习的方法通过提取观点特征，然后训练分类器，最后得到合适的模型来进行观点识别。这种方法割裂了文本中应有的词语间的相互联系，不够灵活全面，无法应用于复杂和灵活的句子。基于规则的观点识别是对语言的表达习惯进行归纳总结，具有较高的准确率，但是其应用范围有限，并且也不适合大规模文本数据。

本研究通过规则判断和机器学习相结合，首先对评论数据集进行观察，根据语言特点，归纳了一些置信度较高的观点判别规则，将明显属于观点句或是属于非观点句的句子提前进行筛选，然后通过机器学习的方法，通过训练分类器将剩余的句子分为观点句和非观点句两类，最终提高观点识别的效率和准确率。

3.3.1 支持向量机

支持向量机（Support Vector Machine，SVM）是 Cortes 和 Vapnik 于 1995 年首先提出的，在小样本、非线性、高维模式识别问题中，相对于其他算法有较大的优势。并且它是建立在统计学习理论的 VC 维理论和结构风险最小理论基础上的，寻求最优间隔分类器（Optimal Margin Classifier）。

支持向量机是一种常用的用于二分类的监督式学习方法，其主要思想可以概括为两点：

（1）它是针对线性可分情况进行分析，对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。

（2）它基于结构风险最小化理论之上，在特征空间中建构最优分割超平面，使得学习器得到全局最优化，并且在整个样本空间的期望风险以某个概率满足一定上界。

例如，针对线性问题，使用 SVM 构建一个简单的线性分类器，用一个简单的二维两类样本分类例子说明，如图 3.2 所示。

图 3.2 中间的直线是一个分类函数，它对 C_1 和 C_2 两类样

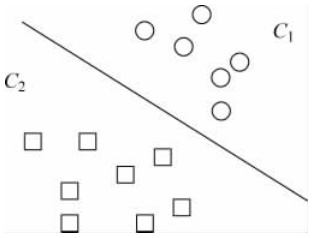

图 3.2 简单的线性分类

本进行划分。这是一个线性函数，在分类过程中，附加一个阈值，通过判断分类函数的执行结果是大于还是小于这个阈值来确定类别。假设这个函数是 $g(x)=wx+b$，设置阈值为 0，若 $g(x_i)>0$，则判别为类别 C_1；若 $g(x_i)<0$，则判别为类别 C_2。此时也等价于给函数 $g(x)$ 附加一个符号函数 sgn()，即 $f(x)=\mathrm{sgn}[g(x)]$ 是真正的判别函数。

对于非线性的情况，把样本从低维度空间映射到高维度空间，将原来的非线性问题转换为线性问题。升维会加大计算的复杂度，甚至引起维度灾难，SVM 通过核函数有效解决了这个问题。因此，针对非线性问题，SVM 的处理方法是选择一个核函数，通过将数据映射到高维空间来解决在原始空间线性不可分的问题，最终找到一个最佳分离超平面对样本进行分类，如图 3.3 所示。

图 3.3 中圆形和方形代表两类样本，H 为分类线，H_1、H_2 分别为各类样本中离分类线最近的样本并且平行于分类线的直线，它们之间的距离叫做分类间隔。其中，超平面记为 $(w,x)+b=0$，设 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，$x\in\mathbf{R}^m$，$y\in\{-1,+1\}$ 为给定样本训练集。其中，n 代表训练样本的数目，m 代表训练样本的维数。通过寻找最优分类面，使得分类间隔

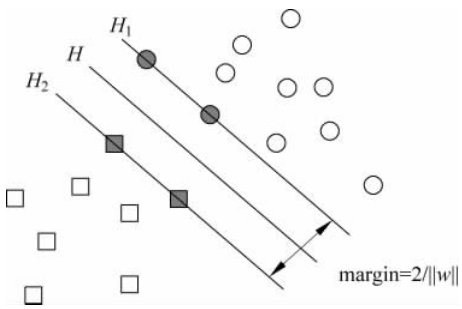

图 3.3 线性可分情况下的最优分类线

最大。

通常选择不同的核函数，可以生成不同的 SVM，常用的核函数有以下四种：

(1) 线性核函数 $K(x,y)=x\cdot y$。

(2) 多项式核函数 $K(x,y)=[(x\cdot y)+1]^d$。

(3) 径向基函数 $K(x,y)=\exp(-|x-y|^2/d^2)$。

(4) 二层神经网络核函数 $K(x,y)=\tanh(a(x\cdot y)+b)$。

在自然语言处理领域中，支持向量机广泛应用于词义消歧、文本自动分类、信息过滤等方面。文献[72]进行对比实验表明支持向量机的分类方法效果最佳，分类精确度最高达到 83%；文献[73]实验表明在训练集规模较大的情况下，使用支持向量机分类方法明显优于其他分类方法。因此，本研究采用支持向量机融合情感特征向量对文本进行文本观点判别。

3.3.2 观点识别过滤规则

本研究通过对评论数据集的观察，设置非观点过滤规则如下：

规则 1：句子中包含超链接，缺少情感词语等直接和观点识

别相关信息，可以判断为非观点句。

规则 2：句子中包含大量数字、乱码、特殊符号等无效信息，可以判断为非观点句。

规则 3：句子中没有与话题相关的评价对象，但是存在和话题无关的评价对象，例如天气、心情、推销等，可以判断为非观点句。

规则 4：仅含有标签，没有实际信息的句子，可以判断为非观点句。

由于电子商务产品评论内容简短且用语不规范，用户表达观点的方式多样。在设置观点过滤规则中，一般认为只要评论包含与电子商务产品质量话题相关的评价对象，那么该条评论就属于话题相关的评论。如果评论中抽取不到任何评价对象，那么默认为话题本身。因此，本研究通过对评论数据集的观察，设置观点过滤规则如下：

规则 1：包含与电子商务产品质量话题相关的评价对象的句子，可以判断为观点句。

例如：#鼠标#不灵敏、不好用！

规则 2：包含网络情感词、情感词且句子中有评价对象，这样的句子可以判断为观点句。

例如：#鼠标#不灵敏、不好用、太可恶了！

规则 3：包含表情符号且句子中有评价对象，这样的句子可以判断为观点句。

例如：#鼠标#［伤心］［哭泣］

文本在使用 SVM 分类器进行观点识别之前，基于规则进行观点识别，流程如图 3.4 所示。

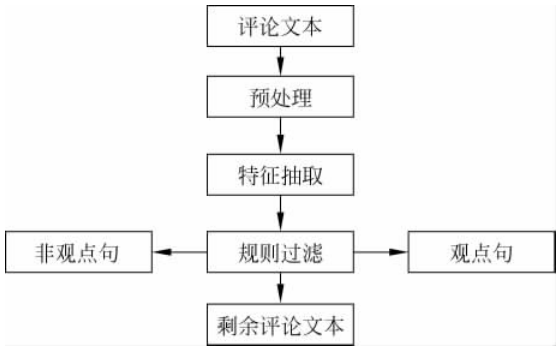

图 3.4 基于规则的初步观点识别

3.3.3 基于规则—SVM 观点识别算法

孙建旺等人研究表明：在中文文本数据集和英文文本数据集中，使用集中典型的文本分类算法，进行性能比较分析，实验结果显示，SVM 算法在精确度方面最高，但是所需的时间开销最大[74]。由于文本分析的话题评论是小规模评论文本，因此不考虑开销时间，最终选择规则与 SVM 结合的观点识别算法。

对于一个文本 d_i 进行预处理后可以将该文本表示为 $d_i=\{t_{i1}, t_{i2}, t_{i3}, \cdots, t_{in}\}$（$t$ 表示特征，i 表示特征的数量）和该文本所属类别 c_i，则文本数据集可表示为 $D=\{(d_1, c_1), (d_2, c_2), (d_3, c_3), \cdots, (d_m, c_m)\}$ 和数量 m，其中 m 表示数据集中文本的数量。

基于规则—SVM 算法描述如下：

输入：文本数据集 $D=\{(d_1, c_1), (d_2, c_2), (d_3, c_3), \cdots, (d_m, c_m)\}$ 和数量 m。

输出：分类结果 $F(d_i)$，$F(d_i)\in C$。

（1）将文本数据集转换成 SVM 模型。

(2) 计算 SVM 模型中每一个特征的信息增益值 IG(T),所得的值保存在一个特征集合中。

(3) 对该特征集合进行排序,并删除小于 0 的值。

(4) 根据新的特征集合重新建立 VSM 模型。

(5) 选择 SVM 分类器的数量。

(6) 对于每一个 SVM 分类器,从新的特征集合中随机生成一个特征子空间样本。

(7) 使用 SVM 分类器对特征子空间样本进行分类。

(8) 结合每个 SVM 分类器结果,最终的输出由多数投票或通过组合后得出。

经过本研究设置的规则进行观点识别之后,剩余评论通过 SVM 分类器进行观点分类。具体步骤:先选取一部分评论文本作为训练样本,并且以句子为单位进行标注。在预处理中使用中科院 ICTCLAS 分词系统对语料进行分词和词性标注,然后进行特征抽取,通过特性项将句子以向量表示,对 SVM 分类器进行训练,得到分类模型,最后对剩余评论文本进行观点分类。具体流程如图 3.5 所示。

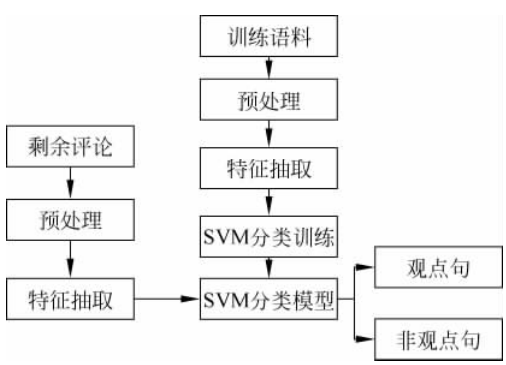

图 3.5 基于 SVM 的观点识别具体流程

3.4 电子商务产品质量评论情感分析实验

3.4.1 实验设置

本研究数据来自于数据堂提供的电子商务产品质量话题型评论文本数据。数据堂是国内专业的科研数据共享服务平台，并且和各大高校、研究机构、企业相互合作，搜集大量的专业和高质量数据，为各种类型用户提供了各种数据需求，将数据价值充分发挥。目前，数据堂的数据库包含语音识别、健康医疗、交通地理、电子商务、社交网络、图像识别、统计年鉴、研发数据等多样化的数据类型，而且还提供更加专业的数据定制服务。

为避免由于单一话题评论的特殊而导致实验的误差本研究选择房价问题、iPhone4 手机两个话题领域的微博评论文本；为增加实验结果的合理性，每个话题评论数量各 3000 条。在下面的各个实验中，以人工计算为标准进行对比，选用四名志愿者对 3000 条数据进行人工标注。

3.4.2 实验的评价指标

为验证本研究提出的方法相比以往情感分析方法的有效性，分别通过人工分类、本研究构建的情感倾向计算模型、现在比较成熟的武汉大学研发的内容挖掘软件中的 ROST_EA[75]情感分析这三种方式分析两类评论数据的情感倾向，并且以人工计算为标准进行对比。采用两种常用的指标，即准确率(Precision)和召回率(Recall)。

准确率指的是测试集中与人工计算结果一致的文本占所有测试集中文本数量的比例，计算公式如下：

$$准确率(\text{Precision}) = \frac{判断正确的类别数目}{判断为该类别的数目} \tag{3.6}$$

召回率指的是测试集中与人工计算结果一致的文本占测试集中所有被人工判定为该类的样本的比例，即被正确预测的样本占所有属于该类样本数量的比例，计算公式如下：

$$召回率(\text{Recall}) = \frac{判断正确的类别数目}{应判断为该类别的数目} \tag{3.7}$$

3.4.3 话题相关领域情感词扩展实验结果

本研究从中文情感词汇本体筛选出基础情感词一共 24 770 个，其中正面情感基础情感词 10 541 个，负面情感基础情感词 10 102 个，中性情感基础情感词 4127 个，构建适用于话题的情感倾向词典，对情感词赋予情感强度。本研究针对两个话题构建领域情感词典，两个话题领域词汇包含的词汇数量：房价问题 218 个，iPhone4 手机领域词汇 271 个。通过人工标注，得出两个话题领域词典的极性准确率如表 3.9 所示。

表 3.9 两个话题领域词典极性准确率

评价指标	话题	
	房价问题	iPhone4 手机
准确率	0.882	0.905

由表 3.9 的结果可以看出：文本构建领域词典的效果较好，具有较高的极性准确率。在构建房价问题领域词典时，准确率相对稍低，主要原因在于房价问题的话题存在一词多义和反

语的情况，导致计算相似度过程中，将负面情感词汇误判为正面情感词汇。

3.4.4 观点识别实验结果

对电子商务产品质量评论文本进行观察，发现评论文本中存在不少和话题无关的非观点句。因此，在进行情感分类之前，进行观点识别。本研究采用规则过滤和 SVM 算法相结合，首先通过设置的过滤规则，将表现明显的非观点句和观点句筛选出来。然后，选取一部分话题评论，作为训练样本，通过训练 SVM 分类器，得到一个合理的分类器，对剩余评论文本进行观点识别。对电子商务产品质量话题评论进行观点识别时，以人工标注为参考。

本研究采用我国台湾大学林智仁教授等开发的易于使用和有效的 SVM 软件包。目前该软件包已经拥有多个版本，包括 Java、MATLAB、C、C♯等，软件包可以被编写的程序直接调用。为了体现采用规则和 SVM 算法相结合方法的有效性和合理性，本研究通过和直接采用 SVM 算法进行对比，具体两种方法的实验结果如表 3.10 所示。

表 3.10 两个话题评论观点识别结果

方法	话题					
	房价问题			iPhone4 手机		
	准确率	召回率	F1 值	准确率	召回率	F1 值
本研究方法	0.831	0.796	0.792	0.872	0.837	0.854
SVM 算法	0.763	0.784	0.773	0.816	0.783	0.799

从表 3.10 可以看出：采用规则过滤和 SVM 算法相结合，确实比直接采用 SVM 算法进行分类的效果好一些。两种方法的各个指标并不是非常高，这和评论文本进行数据预处理中分词和词性标注的准确率有关，不合理的分词会导致有些关键的信息无法被抽取，并且也受特征抽取算法的影响。最后房价问题、iPhone4 手机两个话题经过本研究观点识别之后，得到的观点句数量分别为 1034 条、1115 条。

3.4.5 话题评论情感极性分类结果

文本进行话题评论情感极性分类是建立于观点识别的基础上。通过观点识别之后，对评论文本进行预处理，使用已经构建好的词典，对电子商务产品质量话题情感相关的特征项进行抽取，并获取相应的权值，最后根据电子商务产品质量评论情感计算公式判断其情感极性。为验证本研究构建情感词典和评论情感计算方法的有效性及合理性，设计两个实验进行对比，最后实验结果以人工标注为标准。人工分类结果如表 3.11 所示。

表 3.11 两个话题评论人工极性分类结果

话　　题	正面情感数目	中性情感数目	负面情感数目	总数目
房价问题	137	185	712	1034
iPhone4 手机	415	224	456	1115

在实验一中，实验采用的情感词典包括基础情感词典、程度副词词典、否定词词典、表情符号词典、评价对象词典。评论文本情感判断方法还是采用本研究的情感计算公式。实验结果如表 3.12 所示。

表 3.12 实验一极性分类结果

话　　题	评估指标	正面情感	中性情感	负面情感
房价问题	准确率	0.718	0.686	0.728
	召回率	0.715	0.678	0.736
	F1 值	0.716	0.682	0.732
iPhone4 手机	准确率	0.697	0.661	0.708
	召回率	0.691	0.672	0.712
	F1 值	0.694	0.666	0.710

在实验二中，实验采用的情感词典包括基础情感词典、程度副词词典、否定词词典、表情符号词典、话题领域词典、评价对象词典。评论文本情感判断方法还是采用本研究的情感计算公式。实验结果如表 3.13 所示。

表 3.13 实验二极性分类结果

话　　题	评估指标	正面情感	中性情感	负面情感
房价问题	准确率	0.741	0.709	0.762
	召回率	0.737	0.714	0.743
	F1 值	0.739	0.711	0.752
iPhone4 手机	准确率	0.773	0.701	0.751
	召回率	0.762	0.724	0.758
	F1 值	0.767	0.712	0.754

在实验三中，实验采用的情感词典包括基础情感词典、程度副词词典、否定词词典、表情符号词典、话题领域词典。评论文本情感判断方法还是采用本研究的情感计算公式，但是去掉评价对象影响因素。实验结果如表 3.14 所示。

表 3.14 实验三极性分类结果

话　　题	评估指标	正面情感	中性情感	负面情感
房价问题	准确率	0.657	0.673	0.645
	召回率	0.641	0.682	0.637
	F1 值	0.649	0.677	0.641
iPhone4 手机	准确率	0.674	0.686	0.686
	召回率	0.670	0.673	0.675
	F1 值	0.672	0.679	0.680

(1) 通过实验一和实验二的结果对比,发现电子商务产品质量话题领域词典的加入确实可以提高评论情感分类的准确率。并且话题所包含的领域词汇越多,对评论的情感分类影响就越大,这说明话题领域词语包含评论情感信息,越具专业性的话题包含越多的专业领域词汇,对评论的情感分类具有很大的影响作用。因此,领域情感词典构建的完整度和准确性将会影响评论情感分类的准确率。在负面情感方面,评论数量相对较高,这是因为用户针对某个电子商务产品质量话题评论时负面情感居多。相对而言,负面情感评论分类效果较好。

(2) 通过实验二和实验三的结果对比,针对电子商务产品质量话题,考虑到评价对象对电子商务产品质量评论语句的影响,引入评价对象词典,评论相对于电子商务产品质量话题的情感分类准确率确实有所提高。以往电子商务产品质量评论情感分析中,只考虑和情感相关的特征项,缺乏考虑评价对象。由于电子商务产品质量话题评论往往由多个字句构成,在不同字句中评价的对象常常不同,表达的情感极性也不同,如果还是按照以往不对评价对象加以区分,那么对评论句的情感分类容易造成错误。针对评论中具有多个不同的评价对象的情况,应该考虑评论相对于电子商务产品质量话题的极性。因此,应构建评

价对象词典，将评价对象考虑到评论的情感极性分类中。

3.5 小结

电子商务产品的评论信息对于电子商务产品质量舆情监测具有极大的参考价值。为准确评价电子商务产品质量评论的情感强弱程度，计算电子商务产品质量情感倾向度，本章提出改进以往情感词典中不区分情感词情感强弱的缺陷，构建计算电子商务产品质量评论情感倾向度分类模型，并通过数据验证该模型的科学准确性。最后用支持向量机对互联网上电子商务产品质量评论进行文本情感分类研究。实验表明基于支持向量机的分类器能够有效提高电子商务产品质量评论主题型情感分类准确性，具有分类速度快、健壮性强等特点。