基于财经新闻的金融领域负面情感词典构建研究*

赵又霖[1,2]　林怡妮[2]　孙虹[2]　程丽洁[2]　徐竟楠[2]　陆颖隽[3]

（1. 南京大学 信息管理学院，江苏 南京 210023；
2. 河海大学 商学院，江苏 南京 211100；
3. 武汉大学 信息管理学院，湖北 武汉 430072）

摘　要　金融领域的情感倾向具有其领域特有性，通用情感词典无法准确判断其情感倾向。本文以道琼斯工业平均指数排名前 30 的公司的实时新闻信息为研究对象，设置人工标注规则，采用多种机器学习算法对已标注数据集进行训练，使用特征选择算法抽取特征词语作为负面情感词语，从而构建金融领域负面情感词典。本文以 McDonald Financial Dictionary（麦当劳金融词典）为目标情感词典，通过直接和间接评测两种方式验证其准确性。研究结果表明：本文通过实验获得金融领域 603 个负面的领域情感词，其领域负面情感词典的覆盖率为 82.5%，且对金融领域新闻信息情感识别的准确率达到 92.8%，能够显著提高金融领域负面情感倾向分析的准确率。

关键词　负面情感词典，财经新闻，特征词语，直接评测，间接评测

中图分类号　F832.5

1　引言

信息时代的海量经济信息包括了年报、公告等静态数据以及新闻等动态社交媒体数据，这些信息可以帮助金融机构及投资者在营销、品牌、客户关系管理收益以及监控公众舆论等方面发挥重要作用。投资者的情感倾向、从众心理等非理性因素对投资者的行为将产生直接影响。情感分析是投资者投资决策的重要参考标准，不同的情感极性以及不同程度的情感强度对投资者行为的影响也各有不同。从信息传播的角度来讲，负面信息的传播对个人和社会的稳定发展会造成严重影响[1]，金融领域亦是如此。

利用情感词典进行情感极性判断是目前主流的情感分析方法，词典的全面性和精准性将决定情感分析结果的准确性。现有较为成熟且具有代表性的情感词典包括哈佛大学情感词典、WordNet 情感词典、匹兹堡大学 OpinionFinder 情感词典、HowNet 情感词典、普林斯顿大学情感语义词典以及伊利诺伊大学产品领域评论词典等。由于金融领域的语言具有很强的行业特异性且语言表达方式会随着经济环境的变化而发生改变，故使用通用情感词典对金融领域进行情感分析准确率不高，并且其无法有效而准确地应用到具有专业特色的金融领域。与此同时，现有金融领域情感词典的研究皆是基于期刊、年报等静态数据构建，对实时新闻数据的情感分析较少涉猎。此类静态数据语言较为规范，但词语组成形式不够丰富，随着近几年网络用语的不断更新，依据静态文本数据构建的情感词典难以应对瞬息万变的市场经济分析

* 基金项目：江苏省社科基金青年项目“社会感知数据驱动下的公共卫生事件时空演化研判机制研究”（20TQC001）；中国博士后科学基金特别资助“面向应急管理的时空数据语义模型构建及创新应用机理研究”（2021T140311）；中国博士后科学基金面上项目“环境污染突发事件的时空数据挖掘及协同治理机制研究”（2019M650108）；中央高校基本科研业务费项目“科技创新团队技术多元化对专利颠覆性的影响研究”（项目编号：B230207061）。

通信作者：赵又霖，河海大学商学院副教授，博士生导师，E-mail：sobzyl@hhu.edu.cn。

的需求。McDonald Financial Dictionary[2]作为金融领域的代表性情感词典，是依托静态年报数据构建的。本文通过前期初探发现，该词典对实时媒体新闻数据适配性不高。

动态信息是应急管理实时响应的基础，本文以金融领域实时更新的新闻数据为研究对象，借助机器学习算法构建该领域负面语料库，在负面语料库基础上构建负面情感词典，最后与 McDonald Financial Dictionary 比较并进行直接评测和间接评测来检验该词典的准确率。McDonald Financial Dictionary 是基于 1994 年至 2008 年 50 115 家公司的年报数据构建的金融领域情感词典。该词典作为基础情感词典，可以综合利用市场数据或新闻数据预测股票价格，建立价格预测模型[3, 4]。该词典定义的情感倾向分别为“negative”“positive”“litigious”“uncertainty”“constraining”“superfluous”（“消极”“积极”“诉讼”“不确定性”“约束”“多余”）六个维度。本文主要选择的是负面的情感倾向作为对比的目标情感词典，本文将其称为目标情感词典。本文的研究旨在从具有语言表达特有性的金融领域的实时动态财经新闻数据入手，构建较为完整和精准度较高的领域负面情感词库，以期为后期的预测奠定基础。

2 研究现状

相关研究发现，负面情绪比正面情绪对人的心理和行为影响更大[5]，人们对坏消息的反应比好消息更加敏感[6]。一方面，负面新闻的传播会影响企业股票价格[7, 8]；将《华尔街日报》流行的“与市场并行不悖”专栏与随后的股票回报和交易量联系起来的研究发现：专栏中悲观的词语越多，第二天的回报率就越低[9]；因此许多金融和会计研究机构[9-12]使用文本分析方法研究企业 10-K 报告、报纸文章、新闻稿和投资者留言板的基调和情绪，在与其他金融变量的显著相关性上发现，消极词分类在测量语气方面是有效的。另一方面，投资者在投资时会关注公司的负面新闻，并随之改变投资意愿，如娄岩等[13]以“知乎”数据为例，构建“关注度-满意度”的综合框架，利用该框架分析社交媒体用户对老年科技的关注点和情感倾向趋势，根据用户的负面评论聚类出一些满意度较低的主题，最终根据这些主题确定企业未来的发展和改进方向。从上述研究可以发现：在金融领域，有效地利用负面情感倾向的文本信息既可以帮助企业改善产品，还可以帮投资者避免损失。

目前，情感词典大致可以分为领域独立的情感词典和领域依赖的情感词典两种类型[14, 15]。通用情感词典对特定领域情感分析的方法有两种。其一，直接将通用情感词典作为种子情感词典进行情感分析；其二，以通用情感词典为基础，通过语义相似度计算方式发现并扩充新词。研究证明基于扩展情感字典所提出的情感分析方法具有一定的可行性和准确性[16]，故而有研究者开始通过种子情感词典计算出新的情感词典，从而提高情感分析的准确率。例如，以 HowNet 通用情感词典为种子词典，通过词语与种子词典不同类别中的词语之间的平均相似度判断词语的情感倾向[17]；以基础情感词词典、连词词典等通用情感词典为词典支撑，通过情感特征抽取算法构建出情感词典[18]；以 WordNet 通用情感词典为基准构建的认可度较高的 SentiWordNet 情感词典[19]。

在领域依赖的情感词典研究中，很早就有研究者开始注意特定领域的语料库在数据分析中的独特性。例如，Pang 等[20]在国际著名电影评论网站 Internet Movie Database（网络电影数据库）中影评数据的基础之上，构建了一个可广泛应用于各种粒度的情感分类任务的电影评论语料库。此后，为了提高特定领域情感分析的准确性，国内外很多学者也开始通过独特领域的数据集构建特定领域的情感词典，主要体现在消防[21]、汽车[22]、电影[23]、中文图书[24]、心理健康[25]、电子商务[26]等领域。

就目前情感词典的构建方法来看，较多研究者[27-29]通过计算文本数据中词语与种子情感词典中词语之间的语义相似度来构建情感词典；此外，也有研究人员利用线性规划[30]、图传播算法[31]、融合基于词

典和基于语料的混合算法[32]、上下文信息算法[33]、概率图模型[34]等各种不同类型的算法去构建情感词典。例如，李慧[35]等研究者融合情感词典和机器学习算法对段落级、篇章级文本进行多级情感分类，从而挖掘其中包含的情绪变化。在众多机器学习算法中，不同的算法对数据的表现程度不一样，随着人工智能机器学习算法越来越成熟，情感词典构建开始通过机器学习算法对文本数据进行有监督的分类，故根据目前有监督情感分类方法来看，大多是机器学习方法的分类方法[36]，典型的有朴素贝叶斯算法[37-39]、支持向量机算法[40, 41]、逻辑回归算法[42-44]、最大信息熵[45]等方法。通过研究发现，朴素贝叶斯、逻辑回归、支持向量机算法在文本分类算法中应用较多。例如，李佳儒等[46]使用逻辑回归算法和 TF-IDF 算法对在线文本数据进行情感分类；何跃等[47]使用支持向量机、朴素贝叶斯等算法对微博的博文数据进行三元情感分类。

财经新闻信息所反映的情感倾向往往与市场经济变化趋势有着显著关系，对金融领域的实时动态新闻信息加以挖掘可以帮助进入投资市场的投资者掌握市场动态避免损失，提高预测准确率；从情感词典研究的现状来看，领域依赖性情感词典对于特定领域情感分析的准确率要高于领域独立情感词典，并且相关研究表明金融领域情感词典与通用情感词典存在显著差异，甚至存在表达意义相悖的情况[2]。由此可见，金融领域投资者的需求等客观情况和领域语言情感的特有性共同决定了构建该领域情感词典的必要性与社会价值。

3 基于财经新闻的领域负面情感词典构建

本文以获取的数据源和情感标注的数据集为研究对象，构建金融领域情感词典。实验共分为两部分，第一部分以情感标注的训练集作为研究对象，使用朴素贝叶斯中的伯努利贝叶斯和多项式贝叶斯，逻辑回归算法，支持向量机中的支持向量分类（support vector classification，SVC）、线性支持向量分类算法（linear support vector classification，LinearSVC）和核支持向量分类算法（nu-support vector classification，NuSVC），深度学习中的卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）和长短时记忆网络（long short term memory network，LSTM）这九种分类算法训练最优分类模型，将含有负面情感的数据集训练出来；第二部分基于分类好的负面文本语料库，采用特征选择算法抽取出含有负面情感的特征词，对特征词进行反复人工筛选后形成负面情感词典。

3.1 数据预处理及情感标注

3.1.1 数据来源及预处理

本文选取美国知名财经网站 The Street 作为在线媒体上金融数据集之一，并将美国有线电视新闻网（Cable News Network，CNN）business（财经）频道作为财经数据集补充。本文数据选取的研究对象是道琼斯工业平均指数排名前 30 的公司 2020 年 1 月至 2021 年 6 月的新闻数据，共计 12 032 篇。

英文语句的构成包括单句和复句。单句可以分为疑问句、陈述句、感叹句等简单句型，复句可以分为条件从句、假设从句、状语从句、定语从句等复杂句型，这些句型会改变整个句子的情感倾向，故而在构建金融领域新闻数据集时需要对新闻篇章进行分句。英文以空格为分隔符进行分句，不需要考虑词首、词中、词尾等结构，没有词语位置，每个单词独立构成词语。在分句处理过程中，本文包括但不仅限于以“.！？；”为分句标识，并且在分句过程中删除带有情感倾向的问号、感叹号，保留句号、分号。鉴于英文文本句子结构最少为主语 + 谓语形式，至少由 2 个单词构成，因此，本文在去重后，将每个句子中单词数量小于 2 的句子删除，得到本文的研究数据集共计 562 391 条。

3.1.2 情感标注

为了获得负面文本语料库，本文需要事先标注部分数据集，通过机器学习训练出未标注文本的情感极性。56 万余条数据中随机抽取 10 万条数据，在邀请金融领域专业人员了解标注的标准和方法后，制定人工标注数据集规则，邀请 10 名标注人员（其中包括金融领域专家）进行分工标注，其标注的类别分为消极和其他两大类，分别用情感值–1 和 0 代替。①词性标注：本文在研究中使用部分数据集词性标注后发现句子中量词、代词、助词没有明显的情感倾向，因此在标注过程中只标注可能有情感倾向的动词、名词、副词、形容词以及连词。②情感标注：当句子中没有明确表示情绪的词时，可以联系句子中的其他分句，或者依据句子中的连词，观察句子中是否有转折关系，最后整体把握句子情感，标注出情感极性。最终根据标注原则和标注规则标注出来的句子格式如表 1 所示（示例）。

表 1 人工情感标注数据结果示例

内容	中文翻译	情感标注	负面词	词性标注
A friendly reminder from a new Harvard Medical School study：If you're not sticking to a regular sleep schedule，you could be **hurting** your health	哈佛医学院的一项新研究友好地提醒我们：如果你不坚持规律的睡眠时间表，你可能会损害你的健康	–1	hurting	动词
A growing number of companies are **warning** that the coronavirus will prevent them from meeting sales or profit targets for the first three months of the year	越来越多的公司警告称，冠状病毒将使他们无法实现今年前三个月的销售或利润目标	–1	warning	动词
A growing number of retailers have announced they are temporarily closing stores in an effort to prevent the spread of **coronavirus**	越来越多的零售商宣布，他们将暂时关闭门店，以防止冠状病毒的传播	–1	coronavirus	名词
A house，townhome，or condo is typically the largest asset owned by a consumer	房屋、联排别墅或共管公寓通常是消费者拥有的最大资产	0		
A lot is riding on Americans' willingness to grab their wallets again	很大程度上取决于美国人是否愿意再次掏腰包	0		
A newly discovered asteroid will pass close to Earth today	今天，一颗新发现的小行星将接近地球	0		
A number of mega-banks will report their first-quarter earnings this week，which could give us a more accurate picture of just how **badly** the **Covid-19** crisis has hit the financial sector	许多大型银行将于本周公布第一季度盈利，这可能会让我们更准确地了解新冠危机对金融业的打击有多严重	–1	badly, Covid-19	副词，名词
A surge in spending over China's "golden week" holiday has **highlighted** an encouraging rebound in consumption after the **coronavirus** pandemic ravaged the economy early in the year	中国"黄金周"假期期间的支出激增，突显出今年年初新冠疫情重创经济后，消费出现了令人鼓舞的反弹	–1	highlighted, coronavirus	动词，名词
A Texas couple who have been married for 46 years have both survived Covid-19，while the wife battled cancer	得克萨斯州一对结婚 46 年的夫妇都在新冠中幸存下来，而妻子正在与癌症作斗争	0		
A version of this article first appeared in CNN Business' new Nightcap newsletter	这篇文章的一个版本首次出现在美国有线电视新闻网商业频道（CNN Business）的新 Nightcap 时事通信中	0		
Another federal official is making it clear that despite US President Donald Trump's predictions，there's **hardly** any chance a vaccine will be available to Americans by Election Day	另一位联邦官员明确表示，尽管美国总统唐纳德·特朗普做出了预测，但到选举日，美国人几乎不可能获得疫苗	–1	hardly	副词
Alvin Yau is **exhausted**	艾文·邱已筋疲力尽	–1	exhausted	形容词

3.2 负面情感语料库的构建

3.2.1 文本情感分类器构建

本文将情感标注的训练集作为研究对象，使用分类算法训练最优分类模型，构建金融领域负面文本语料库，因此文本情感分类模型的选取结果直接影响构建的负面文本语料库的准确性和全面性。本文只针对文本二元分类问题，大量研究指出支持向量机在二元分类任务中具有较高的效率和准确性[48-50]，但由于金融领域文本语言具有领域专有性和独特性，因此本文为选择最优的模型算法，首先对朴素贝叶斯、逻辑回归、支持向量机以及深度学习四大类方法进行评估。实验对训练好的模型进行评估，目的在于选取更优的算法进行训练比较，进而训练出准确率较高的负面文本语料库。

1. 基于朴素贝叶斯算法的情感分类

（1）伯努利贝叶斯[51]。本文实验数据集被分为“消极”或“其他”，因此通过二元形式可表示为

$$w=\{w_1,w_2,\cdots,w_n\},w_m\in\{0,1\} \tag{1}$$

其中，$w_m=1$为该词 m 在消极文档 w 中出现，反之则未出现；n 为文档总数。

伯努利贝叶斯算法中，预测文档 w 的类别概率 $c(d)$ 的公式如下：

$$c(d)=\operatorname{argmax}\frac{\sum_{i=1}^{n}\delta(c_i,c)}{n}\prod_{i=1}^{n}\left\{w_n\frac{\sum_{i=1}^{n}w_i\delta(c_i,c)+1}{\sum_{i=1}^{n}\delta(c_i,c)+2}+(1-w_i)\left[1-\frac{\sum_{i=1}^{n}w_i\delta(c_i,c)+1}{\sum_{i=1}^{n}\delta(c_i,c)+2}\right]\right\} \tag{2}$$

其中，$\sum_{i=1}^{n}\delta(c_i,c)$ 为属于类别 c 的文档总数。

（2）多项式贝叶斯。在待测文档给定的情况下，其预测公式如下：

$$c(d)=\operatorname{argmax}\left[\log_2\frac{\sum_{i=1}^{n}\delta(c_i,c)+1}{n+1}+\sum_{i=1}^{n}\log_2\frac{\sum_{i=1}^{n}f_\mu\delta(c_i,c)+1}{\sum_{i=1}^{m}\sum_{i=1}^{n}f_\mu\delta(c_i,c)+m}\right] \tag{3}$$

其中，f_μ 为第 i 句新闻数据中第 i 个单词的频率，μ 取值为[1, n]；c_i 为第 i 篇新闻数据的类标记；m 为属性个数。

本文将已标注的数据集作为训练集，得到两种情感分类算法的准确率有细微差异。其中，多项式贝叶斯算法的准确率为 81.5%，伯努利贝叶斯算法的准确率为 81.1%；从准确率来看，前者的准确率略高于后者。

2. 基于逻辑回归算法的情感分类

本文使用逻辑回归的主要任务是通过二元分类，根据输入的已标注数据集计算出未标注数据集归属的每种类别的概率[52]。在本文中每个对象的标签是一个二维向量，如果预测对象属于 i 类，则标签向量的第 i 位的值为 1，其余位的值为 0。也就是说，如果预测对象属于消极类，则标签为 y = (1, 0)，如果预测对象为其他类，则标签为 y = (0, 1)。为了能够根据标注好的句子 x 预测其他句子 Pr(y = 1)的概率，需要寻找一个连续函数，既能表达 x 与概率 Pr(y = 1)之间的关系，也能保证特征变动时对应的函数值不超过[0, 1]。本文使用 Sigmoid 函数满足这种要求，其函数表达式如下：

$$h_w(x)=\mathrm{Sigmoid}\left(\langle w,x\rangle\right)=\frac{1}{1+\mathrm{e}^{-\langle w,x\rangle}} \tag{4}$$

其中，$h_w(x)$ 为逻辑回归函数图像的纵坐标，特征值组 $x\in R^n$；$\langle w,x\rangle=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$，$x$ 为数据向量集$\{x_1,x_2,\cdots,x_n\}$；β 为逻辑回归模型的参数。

本文的阈值设置为 0.5，如果 $p(y=1|x,w)$ 大于 0.5，则可以判定该数据的分类为消极，否则为其他。使用逻辑回归算法进行情感分类的准确率为 81.2%，与本文的数据集较为适配。

3. 基于支持向量机算法的情感分类

本文主要对未标注数据集进行二元分类处理，即大小为 n 的训练样本集由二类别组成$\{(x_i,y_i),i=1,2,3,\cdots,n\}$，如果 $x_i\in R^n$ 预测结果为消极类，则标记为负面（$y_i=1$），如果预测结果不为负面，则标记为负（$y_i=-1$）。为选择最优算法，本文将使用支持向量分类、线性支持向量分类以及核支持向量分类算法对数据集进行训练[53]。

（1）支持向量分类算法。设训练数据集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$x$ 为平面上一点，直线 $l=\langle w,x\rangle+b$ 是平面上的分离直线，则对于任意$1\leqslant i\leqslant m$，均有 $y_m=\mathrm{Sign}\langle w,x\rangle+b$。两者分类间隔达到最大的公式如下：

$$\max\delta(w,b)=\max\left\{\min\frac{y_m\left(\langle w,x\rangle+b\right)}{\|w\|}\right\} \tag{5}$$

其中，$\dfrac{y_m\left(\langle w,x\rangle+b\right)}{\|w\|}$ 为点 x 到直线 l 的距离；$\delta(w,b)$ 为 l 与训练集 S 的间隔。

（2）线性支持向量分类算法。给定训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}\in(X\times Y)^t,x_i\in R^n,y_i\in Y=\{-1,1\},i=1,2,\cdots,n$，$R^n$ 为输入的特征值组，线性支持向量分类算法主要通过求解最优方案 α：$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$，使得函数 L 达到最大。当函数 L 达到最大时，计算 $w=\sum_{i=1}^{n}\alpha_iy_ix_i$，选取 α 的某一分量 $0<\alpha_j<C$，求出 $b=y_j-\sum_{i=1}^{n}y_i\alpha_ix_ix_j$ 的值，最终构建决策函数如下。

$$g(x)=\mathrm{sign}(wx+b) \tag{6}$$

（3）核支持向量分类算法。上述两种方法依赖于积极数据和消极数据之间存在分离平面，也就是说，一旦两种分类之间不存在分离平面，那么本文的二元分类问题就不能达到最优效果，因此本文在使用中加入核方法，将非线性转换成线性问题，通过核函数 $K(x_i\times x_j)$ 将一个空间的样本数据映射到另一个特征空间。核支持向量分类算法主要通过合适的核函数 $K(x_i\times x_j)$ 以及违反约束的参数 C，求解最优方案 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$，使得函数 L 达到最大。当函数 L 达到最大时，选取 α 的某一分量 $0<\alpha_j<C$，求出 $\beta_0=y_j-\sum_{i=1}^{n}y_i\alpha_iK(x_i\times x_j)$ 的值，最终构建决策函数为

$$g(x)=\mathrm{sign}\left[\sum_{i=1}^{n}\alpha_iy_iK(x_i\times x_j)+\beta_0\right] \tag{7}$$

实验通过使用支持向量分类算法、线性支持向量分类算法以及核支持向量分类算法进行情感分类，三种算法的准确率相差较大。其中，线性支持向量分类算法的准确率为 79.2%，核支持向量分类算法的准确率为 86.3%，支持向量分类算法的准确率为 81.5%。

4. 基于深度学习算法的情感分类

通过深度学习算法进行情感分类，需要将文本内容转化为数值表示，并提取出与情感有关的特征。本文采用了 CNN、RNN 和 LSTM 三种经典的深度学习方法对文本数据进行情感准确率的测量。根据 8：1：1 的比例划分训练集、测试集和验证集，基于三种算法进行情感分类，得到的结果是 CNN 分类准确率为 71.0%，RNN 的准确率为 50.3%，LSTM 方法的准确率为 68.6%。

3.2.2 基于机器学习算法的情感极性识别准确率评估

本文在训练数据集的过程中，分别使用上述提到的四大分类算法进行训练。在模型构建阶段中，会留出一部分有标记的训练数据，本文选取已标注数据集的 20%，然后对留出的这部分数据集使用不同参数值评估模型从而得到算法的准确率（其中剩余 80%的数据集与最终用来预测的测试集不同）。在模型选择阶段结束后，利用优化后的模型对未标记的样本进行预测。在机器学习分类训练集过程中，评价分类算法的效果可以通过模型的开销（速度和内存）进行评价，但是分类结果的准确率是评价各种机器学习算法优劣的最直接依据。9 种算法的准确率如图 1 所示，支持向量机中的分类准确率要高于朴素贝叶斯、逻辑回归和深度学习算法；其中，朴素贝叶斯分类算法中，伯努利贝叶斯方法的准确率为 81.1%，多项式贝叶斯的准确率为 81.5%；逻辑回归算法为 81.2%；支持向量机分类算法中，线性支持向量分类算法的准确率为 79.2%，核支持向量分类算法为 86.3%，支持向量分类算法的准确率为 81.5%；深度学习方法中，CNN 分类准确率为 71.0%，RNN 的准确率为 50.3%，LSTM 的准确率为 68.6%。

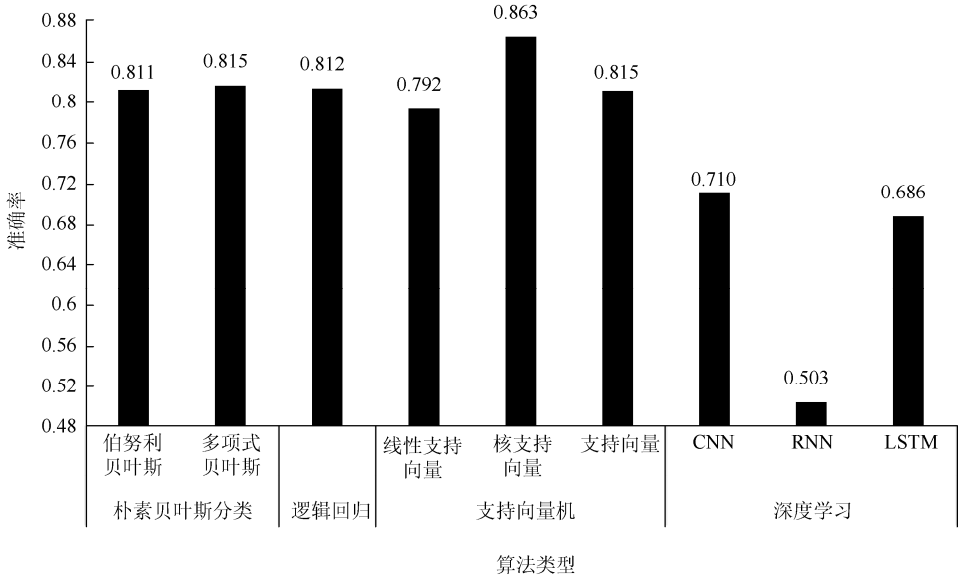

图 1 9 种算法的情感识别准确率比较

因此，从准确率来看，基于支持向量机的算法在二元分类效果上优于朴素贝叶斯、逻辑回归以及深度学习方法。支持向量机是通过有监督的机器学习方式对数据进行二元分类，主要是对已标注的数据集求解出语句分类的最大边距超平面[53]。在实际处理过程中，会出现很多满足这个条件的超平面，因此支持向量机的目标就是构造一个能正确划分已标注好的数据集，使即将分离的两类数据的超平面拥有最大间隔，超平面不仅可以完美地区分训练集，而且能对测试数据集有较好的分类预测效果。本文在使用支

持向量机算法情感分类时，核支持向量分类算法准确率为 86.3%，优于支持向量分类以及线性支持向量分类算法，故本文选择核支持向量分类器进行训练，将未标记的数据集输入核支持向量模型内，最终获得负面文本数据集共 187 117 条数据。

为了验证核支持向量分类器构建的基于财经新闻的领域负面语料库可靠性，选取 3 名参与文本情感分类标注的人员，从构建的负面新闻语料中随机抽取 10000 条文本进行标注，对比两者的结果，进行信度检验。通过 SPSS 可信度分析，Cronbach's α 系数（克龙巴赫 α 系数）为 0.886，达到科学研究的信度。说明本文选取核支持向量分类器构建的负面数据集具有较高的准确性。然后，对获得的负面数据集进行预处理，去除掉停用词并进行词形还原，最终构建成完整、准确的负面语料库。

3.3 金融领域负面情感词典的构建

本文从负面语料库中抽取出具有负面情感倾向的词语构建负面情感词典。通过研究发现，负面语料库中的高频词如“economic”“stock”（“经济”“股票”）等词在语料库中出现的频率较高但不具有情感倾向，因此仅依靠词频来筛选负面情感词语准确率不高。本文考虑到构建的负面新闻语料库中每条句子均有一个情感倾向且每个句子的情感倾向是由句子的基础情感词语确定的，在选择关键词抽取的特征算法时发现 TF-IDF 算法的思想是基于每条句子和每个语料库抽取出最有意义的词语，与单纯依靠词频确定负面情感词语的方法相比该算法可以过滤出频率较高的非情感词语[54]，故选取 TF-IDF 算法对负面的数据集进行特征词语抽取。词频（term frequency，TF）表示句子中的词语在文档中出现的频率。在实际处理过程中，会出现同一词语在长句子里面出现的次数比短句子多的情况，为避免计算结果偏向长文档，本文对词频进行归一化处理，如式（8）所示：

$$\mathrm{TF}_{t,d}=\frac{f_{t,d}}{n_d} \tag{8}$$

其中，$\mathrm{TF}_{t,d}$ 为词语 t 在文档 d 中的词频；$f_{t,d}$ 为文档 d 中，词语 t 出现的数量；n_d 为文档 d 中所有词语的总数量。在实际处理过程中，会出现同一词语在长句子里面出现的次数比短句子多的情况，为避免计算结果偏向长文档，本文对词频进行归一化处理。

逆文档频率（inverse document frequency，IDF）衡量的是词语在文档中的重要性，表示的是词语在数据集中出现的次数。逆文档频率主要是为了排除一些助词、介词、虚词等无意义的词，这些词没有代表性，重要性不高。在实际处理过程中，由于某些词语可能不在文档集中，因此逆文档频率数归零。为避免这种情况，本文将分母调整为 $1+\mathrm{df}_t$，如式（9）所示：

$$\mathrm{IDF}_t=\lg\frac{N}{1+\mathrm{df}_t} \tag{9}$$

其中，IDF_t 为词语 t 的逆文档频率；N 为所有文档数量；df_t 为文档频率，表示包含词语 t 的所有文档的数量。

最终，TF-IDF 算法将词频和逆文档频率结果结合起来，最终表达方式如式（10）所示，其中，$\mathrm{TF\text{-}IDF}_{t,d}$ 为词语的词频-逆文档频率。

$$\mathrm{TF\text{-}IDF}_{t,d}=\mathrm{TF}_{t,d}\times\mathrm{IDF}_t \tag{10}$$

本文为了验证 TF-IDF 算法在抽取负面情感词语方面的有效性，从构建的负面语料库中随机抽取 1000 条数据，以人工标注的负面情感词为基础，从负面情感词语的准确率、召回率和 F1 值三个方面对比基于词频和基于 TF-IDF 算法在抽取负面情感词语方面的差异性。其中，假设预测为负面词语，实际

为负面词语的数量为 TP；预测为负面词语，实际为非负面词语的数量为 FP；预测为非负面词语，实际为负面词语的数量为 FN；词语的总数为 M，那么准确率(P) = TP/M；召回率(R) = TP/(TP + FN)；F1 值(F1) = 2 × (P × R/(P + R))。另外，在验证不同方法抽取负面情感词语的性能之前，考虑选取负面语料库中 Top$N$ 个词语作为负面情感词语时，N 的大小会影响到负面情感词典的准确性和全面性，因此将 N 的阈值分别设为 10%、20%、30%、40%，对比同一种方法在不同候选情感词语选取规模下负面情感词典构建效果。

从图 2 可以看出，比较 TF-IDF 算法和基于词频情感词语抽取方法，显然本文使用 TF-IDF 算法的各项指标均优于基于词频的情感词语抽取方法。对于 TF-IDF 算法识别负面情感词语的准确率随着取词数量的变大而减少的情况，当候选情感词语规模变大时，某些非负面倾向的词语也被抽取为候选情感词语，导致准确率下降，因此准确率随阈值的升高而下降是合理的。F1 值和召回率没有随着阈值的变化而明显变化，说明本文基于负面语料库利用 TF-IDF 算法能够较好地判断候选情感词语的情感极性。另外，从阈值变化来看，当利用 TF-IDF 算法选取特征值排名前 30%的词语作为候选情感词语时召回率最大。

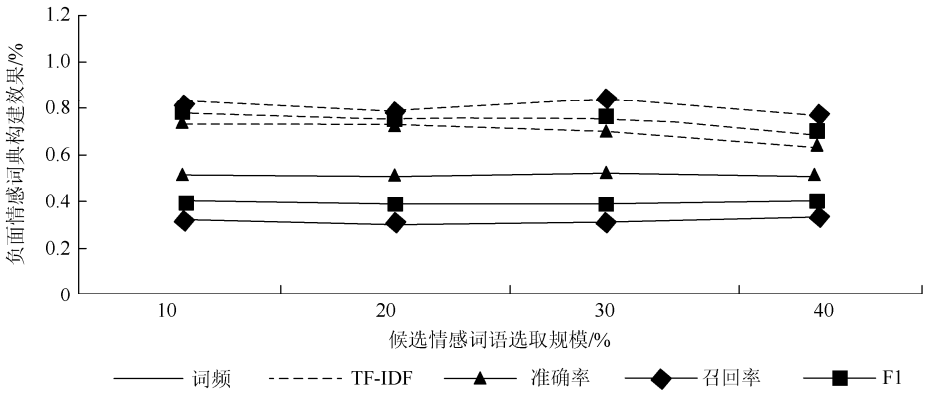

图 2　不同阈值下 TF-IDF 算法和基于词频的方法构建负面情感词典的效果

在统计金融领域情感词典时，虽然在数据处理部分清除了大量的特殊符号和去除停用词，但是运用特征算法计算出来的特征关键词仍然有干扰词语，因此本文在使用特征算法对负面文本语料库进行特征词提取后（前 30%的关键词），人工将对情感分析没有贡献的词语从词典中剔除，从而保证剩余词语对情感分析皆有贡献。人工筛选的情感词语有一定的筛选标准：①在金融学不同文本中具有不同意义的词语。具有不同意义的词语对文本的情感分析具有一定的干扰性，有的词语在一定语境中可以理解为其他意思，这会大大降低金融学文本情感分析的准确性。②对于金融领域不敏感的词语。在分析财经类新闻时，在新闻语句中会存在一些不敏感的词语，这些词语使用率低，在统计词典的过程中予以剔除。此外，本文还将删除频次小于 3 的词语。

通过人工反复筛选语料库，提取出基于财经新闻的金融领域负面情感词典特有的词语，匹配 McDonald Financial Dictionary 词语，剔除部分情感程度低的词语和中性词，最终统计获得 603 个情感词，见附录。

4　金融领域负面情感词典评测与结果分析

4.1　直接评测

本文采取直接评测[25]的方法评测构建的金融领域负面情感词典对 McDonald Financial Dictionary 的

覆盖率。覆盖率指找出金融领域负面情感词典中与目标情感词典重复的词语，对平均准确率进行测算，找出金融领域负面情感词典与目标情感词典排序相同的词语。由于 McDonald Financial Dictionary 负面情感词典中的词语未进行词性还原，因此目标情感词典经过词性还原后发现目标情感词典中共有 412 个词语，其中 340 个数据是重合的。另外，由于目标情感词典在公开发布时是以字母顺序排列而成，因此本文构建的结果也按照字母大小来排序以计算平均准确率，直接评测的覆盖率和平均准确率如表 2 所示。

表 2 直接评测结果

指标	金融领域负面情感词典
词语数量	603
覆盖率	0.825
平均准确率	0.793

从表 2 可以看出，本文构建的负面情感词典的词语数量为 603，远高于目标情感词典中的负面情感词，覆盖率为 82.5%，平均准确率为 79.3%。

4.2 间接评测

间接评测指通过实际数据评估本文构建的负面情感词典的准确率。具体流程如图 3 所示。本文的实验结合 Apple（苹果）公司股价变化的情感分类评估对词典进行间接评测。

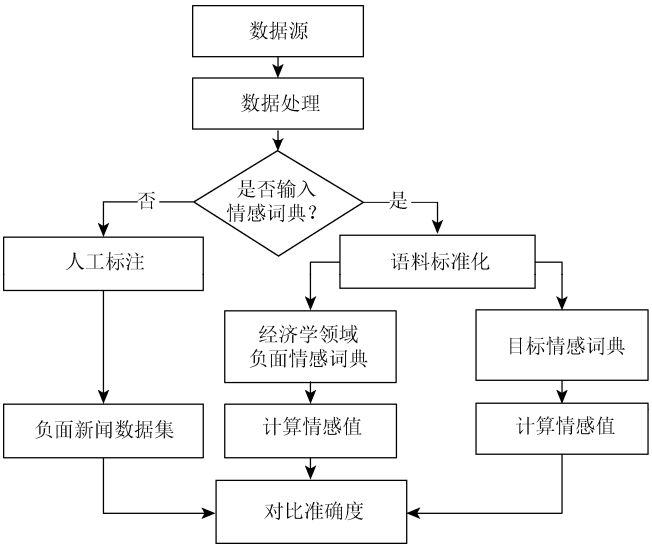

图 3 间接评测流程图

4.2.1 情感识别准确性实验

本文为了验证以前段时间为基准构建的情感词典的实践指导意义，选取后段时间较为活跃的 Apple 公司 2021 年 11 月 1 日至 2021 年 12 月 31 日财经新闻信息为研究对象，共计 169 篇新闻。经过人工标注后获得 6438 条负面数据，分别使用目标情感词典 McDonald Financial Dictionary 与本文构建的金融领域负面情感词典，将情感分析的结果与标注的负面数据的结果进行对比。其中，准确率 Ac 的表达式如式（11）所示：

$$\mathrm{Ac}=\frac{\text{预测结果为负面的数据}}{6438} \tag{11}$$

最终两者的情感分析准确率对比结果如图 4 所示。

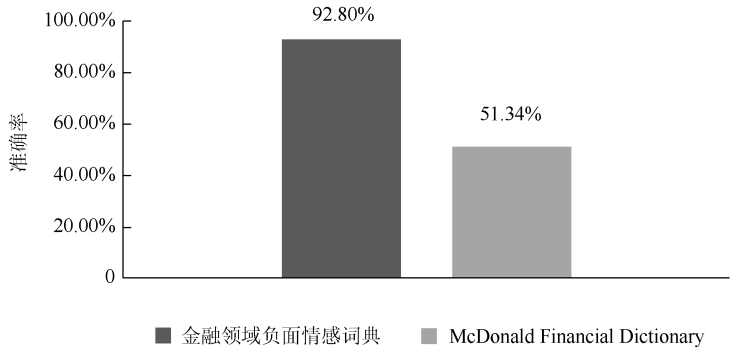

图 4　情感分析准确率评测结果

可以看出,本文构建的金融领域负面情感词典的准确率为 92.80%,目标情感词典 McDonald Financial Dictionary 的准确率为 51.34%。因此，使用本文构建的负面情感词典可以显著地提高情感分析效果。

4.2.2　股价变动及情感变化趋势一致性实验

本部分评测选取 Apple 公司 2021 年 9 月 1 日至 12 月 31 日的新闻信息数据，共计 256 篇新闻数据，得到数据集 12 526 条。结合 Apple 公司的股票价格变化发现，该公司的股票在此时间段价格波动较大，如图 5 所示，其中 9～10 月 Apple 公司股票价格整体处于下降趋势，11～12 月底公司股票价格呈现大幅上涨的趋势，说明在此期间 Apple 公司的经济变化明显，因此将财经新闻数据按照股票的波动情况将新闻数据分成 9 月 1 日至 10 月 10 日和 10 月 11 日至 12 月 31 日两部分，分别对应为股票价格下跌期和上涨期，从而对比分析股票变化不同时期财经新闻文本情感变化的差异。

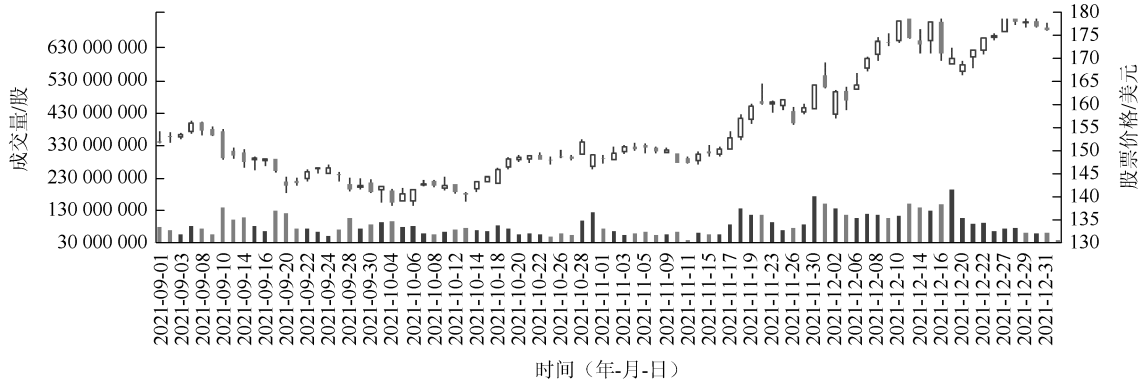

图 5　Apple 公司 2021 年 9～12 月股票价格变化图

利用本文构建的情感词典对该阶段 Apple 公司的新闻数据进行时序分析，分析新闻情感变化与公司股价走势间的一致性关系，从而进一步评估情感词典的负面情感识别效果。其中，具有负面倾向的新闻信息所占比例最大，其次是中性情感倾向的新闻信息，正面倾向的新闻信息所占比例最小，总体来说新闻数据集的情感倾向为负面情感倾向。如图 6 所示，从 9～10 月与 11～12 月不同情感倾向的新闻数据

量对比来看，在股票价格下跌期和上涨期间，虽然新闻信息的情感倾向都是负面的，但不同情感倾向的新闻数据量所占比例有所差异，在 9～10 月，具有负面情感倾向的新闻数据量所占比例为 60.71%，中性情感倾向的新闻数据量所占比例为 28.97%，正面情感倾向的新闻数据量所占比例为 10.32%；11～12 月，具有负面情感倾向的新闻数据量所占比例为 51.31%，中性情感倾向的新闻数据量所占比例为 35.15%，正面情感倾向的新闻数据量所占比例为 13.54%，相对于股票价格走势下跌而言，股票价格上涨期间负面情感倾向的新闻信息所占比例有所下降，中性和正面情感倾向的新闻信息所占比重上升，研究结果表明，财经新闻文本当负面情感倾向的新闻信息所占比例下降时，股票价格走势是上升的，事件变化相符。

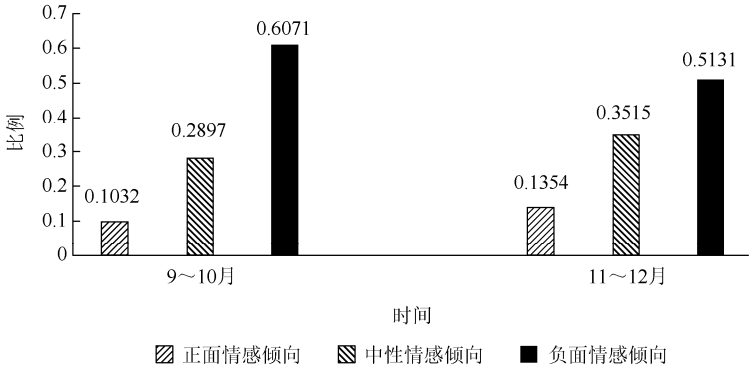

图 6 Apple 公司 9～12 月相关新闻文本的情感分析图

4.3 评测结果的原因分析与启示

4.3.1 原因分析

通过直接评测和间接评测，均发现本文构建的情感词典对情感分析的效果高于目标情感词典，其原因可能如下。

（1）构建情感词典的语料库选取方面。①本文的研究对象为 2020 年 1 月至 2021 年 6 月的实时新闻信息，多数为新冠疫情相关话题，而目标情感词典选取的是 1994 年至 2008 年间 50 115 家公司的年报数据。构建的情感词典从时间上来看，1994～2008 年的经济环境与现在的经济环境完全不同，故目标情感词典对现阶段财经新闻数据的情感分析有一定的局限性。②从数据类型上来看，目标情感词典选取的是公司年报数据，年报数据较为单一，且随着时代的不断变化，20 多年前的经济学语言与现如今的经济学语言有所不同，故目标情感词典对现阶段财经新闻信息的情感倾向判别的准确率不高。③从数据背景上看，本文所选取的财经新闻数据是在新冠疫情背景下的新闻文本数据，该阶段财经新闻报道皆与新冠疫情事件关联。quarantine、lockdown、coronavirus、COVID-19（检疫、封锁、冠状病毒、新型冠状病毒感染）等词都是新冠疫情背景下产生的新词，在目标情感词典中并不包含此类情感词。因此，用本文构建的情感词典对新冠疫情背景下的财经新闻文本进行情感分析的准确性要高于目标情感词典。④本文选取动态变化的实时新闻数据作为研究对象，为动态变化的情感分析奠定了坚实的数据基础。

（2）人工标注的数据集方面。本文数据集多为已经分类好的训练集，故在构建该领域负面语料库时，需要进行人工标注。在标注数据集时，本文遵循科学性、独立性、可操作性、全面性等四个原则，在充分咨询金融领域专家的前提下设置人工标注原则，主要从英文分词、词性标注、情感标注三个步骤执行；

同时邀请 10 位金融领域专家历时一个半月共标注 10 万条数据。从人工标注过程的严谨程度来看，在建立标注规则、邀请标注人员和标注数据量三方面皆保证了人工标注结果的准确性。

（3）负面语料库构建方面。本文采用朴素贝叶斯、逻辑回归、支持向量机以及深度学习四大类方法对数据集训练进行充分的比较，实验结果从定量的角度也充分证明了支持向量机方法在二元分类中的优势，此点也正好印证本文在构建负面领域情感词典中所定义的“消极”和“其他”这一规则。因此，本文在负面语料库构建方法的选取上也采取了科学合理的机器学习方法，进而提高语料库的准确性。

4.3.2 启示

财经新闻的领域情感词典的研究，是为有效识别动态新闻信息联动所产生的负面效应。

（1）政府应及时掌握市场情感倾向变化并调整相关经济政策。2020 年新冠疫情的扩散使得经济环境受到持续的影响。经济市场的低迷影响人们的生产生活，如生活用品、医疗用品的供不应求，非必需商品被大量囤积。利用本文构建的金融领域情感词典分析市场新闻信息，可以准确判断舆情趋势，掌握市场波动情况，政府及相关部门可以调整应对的经济行为及政策，同时可以增强企业活力、稳定市场波动、平复人们的恐惧心理、维护社会稳定和谐。

（2）企业应及时关注网络信息并拓宽运营方式。随着理性人到非理性人的不断发展，投资者在投资过程中一旦发现投资对象的负面新闻，将改变投资行为，从而导致公司股票价格下降，造成企业损失。本文研究对象恰好处于 2020 年新冠疫情扩散期间，此研究结果正好可以用来指导负面情绪下如何调整产品运营策略，使企业能够复工复产，保持经济平稳发展。

（3）个人应根据信息的情感倾向变化调整投资方式。在“互联网 +”的大环境下，人们获取信息的渠道更多元化，投资行为更为理性，决策更加高效。对动态实时信息的情感倾向变化进行分析挖掘，更有助于投资者规避风险，使得投资更为合理、科学。

5 结语

研究发现，目前金融领域的情感分析大多基于通用词典，未考虑到金融领域语言的独特性，针对金融特定领域的研究尚较为空白，本文则是对此领域的探索。①从研究视角来看，针对金融领域语言的独特性，本文从实时财经新闻入手构建金融领域情感词典，使得该词典本身就具有时效性和可观测性。②从词典构建方法来看，综合机器学习算法和特征选择算法完成该领域负面情感词典构建，使得研究范式更加科学化、结果更具实用性。③从词典评估的方法来看，本文结合直接评测和间接评测。特别是以 Apple 公司在往后时间段（2021 年 11～12 月）的财经新闻为评测对象，直观地从动态变化的视角验证基于前段时间（2020 年 1 月～2021 年 6 月）的财经新闻数据构建的情感词典的有效性，并验证 Apple 公司 2021 年 9～12 月的新闻信息情感倾向变化趋势与股票价格变化趋势一致性关系，使得本文更具有实践指导意义。

然而，本文仍需做进一步的完善。①数据源方面，在后续的研究中可以选取多元化的数据，不仅是文本数据，可以考虑从更细粒度的维度定义数据源的情感极性。②在负面情感词语识别方面，本文结合特征算法以词频和逆文档频率相结合的方式构建负面情感词典，未来研究可考虑将负面语料库中后续候选情感词的语义信息加入情感词的情感倾向识别中。③本文主要是通过动态财经新闻信息构建负面情感词典，在后续的研究中可以尝试构建积极情感词典、程度副词情感词典、连词词典等金融领域情感词典，进一步填补该领域词典的空白。

参 考 文 献

[1] 邓利平. 论负面新闻的特征及传播功能[J]. 新闻界，2002，（1）：14-16.

[2] Loughran T，McDonald B. When is a liability not a liability？textual analysis，dictionaries，and 10-ks[J]. The Journal of Finance，2011，66（1）：35-65.

[3] Li X D，Wu P J，Wang W P. Incorporating stock prices and news sentiments for stock market prediction：a case of Hong Kong[J]. Information Processing & Management，2020，57（5）：102212.

[4] Li X D，Xie H R，Chen L，et al. News impact on stock price return via sentiment analysis[J]. Knowledge-Based Systems，2014，69：14-23.

[5] Baumeister R F，Bratslavsky E，Finkenauer C，et al. Bad is stronger than good[J]. Review of General Psychology，2001，5（4）：323-370.

[6] Nguyen C P，Schinckus C，Hong Nguyen T V. Google search and stock returns in emerging markets[J]. Borsa Istanbul Review，2019，19（4）：288-296.

[7] Hanlon M，Slemrod J. What does tax aggressiveness signal？Evidence from stock price reactions to news about tax shelter involvement[J]. Journal of Public Economics，2009，93（1/2）：126-141.

[8] Carvalho C，Klagge N，Moench E. The persistent effects of a false news shock[J]. Journal of Empirical Finance，2011，18（4）：597-615.

[9] Tetlock P C. Giving content to investor sentiment：the role of media in the stock market[J]. The Journal of Finance，2007，62（3）：1139-1168.

[10] Antweiler W，Frank M Z. Is all that talk just noise？The information content of Internet stock message boards[J]. The Journal of Finance，2004，59（3）：1259-1294.

[11] Li F. Annual report readability，current earnings，and earnings persistence[J]. Journal of Accounting and Economics，2008，45（2/3）：221-247.

[12] Tetlock P C，Saar-Tsechansky M，Macskassy S. More than words：quantifying language to measure firms' fundamentals[J]. The Journal of Finance，2008，63（3）：1437-1467.

[13] 娄岩，杨嘉林，黄鲁成，等. 基于网络问答社区的老年科技公众关注热点及情感分析：以“知乎”为例[J]. 情报杂志，2020，39（3）：115-122.

[14] Cardie C. Sentiment analysis and opinion mining[J]. Computational Linguistics，2014，40（2）：511-513.

[15] 王科，夏睿. 情感词典自动构建方法综述[J]. 自动化学报，2016，42（4）：495-511.

[16] Xu G X，Yu Z H，Yao H S，et al. Chinese text sentiment analysis based on extended sentiment dictionary[J]. IEEE Access，2019，7：43749-43762.

[17] 柳位平，朱艳辉，栗春亮，等. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29（10）：2875-2877.

[18] 朱艳辉，栗春亮，徐叶强，等. 一种基于多重词典的中文文本情感特征抽取方法[J]. 湖南工业大学学报，2011，25（2）：42-46.

[19] Baccianella S，Esuli A，Sebastiani F. SentiWordNet 3.0：an enhanced lexical resource for sentiment analysis and opinion mining[C]. Valletta：The Seventh International Conference on Language Resources and Evaluation，2010.

[20] Pang B，Lee L，Vaithyanathan S. Thumbs up？Sentiment Classification using Machine Learning Techniques[C]// Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Stroudsburg：Association for Computational Linguistics，2002：79-86.

[21] 张鹏，王桂玲，徐学辉. 云计算环境下适于工作流的数据布局方法[J]. 计算机研究与发展，2013，50（3）：636-647.

[22] 蒋翠清，郭铁博，刘尧. 基于中文社交媒体文本的领域情感词典构建方法研究[J]. 数据分析与知识发现，2019，3（2）：98-107.

[23] Wang Q Y，Zhu G L，Zhang S X，et al. Extending emotional lexicon for improving the classification accuracy of Chinese

film reviews[J]. Connection Science，2020，33（11）：1-20.
[24] 郭顺利，张向先. 面向中文图书评论的情感词典构建方法研究[J]. 现代图书情报技术，2016，（2）：67-74.
[25] 王东彬. 心理健康词典的自动构建研究：以自杀线索词典为例[D]. 南昌：江西财经大学，2019.
[26] Wang H T，Ren J L. Research on the validity of online commodity reviews based on Word2Vec[C]. Xiamen：The International Conference on Information Technology and Electrical Engineering 2018，2018.
[27] Blitzer J，Dredze M，Pereira F，et al. Biographies，bollywood，boom-boxes and blenders：domain adaptation for sentiment classification[C]. Prague：The 45th Annual Meeting of the Association of Computational Linguistics，2007.
[28] Saif H，Fernandez M，Kastler L，et al. Sentiment lexicon adaptation with context and semantics for the social web[J]. Semantic Web，2017，8（5）：643-665.
[29] Deng S Y，Sinha A P，Zhao H M. Adapting sentiment lexicons to domain-specific social media texts[J]. Decision Support Systems，2017，94：65-76.
[30] Choi Y，Cardie C. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification[C]. Morristown：The 2009 Conference on Empirical Methods in Natural Language Processing，2009.
[31] Feng S，Kang J S，Kuznetsova P，et al. Connotation lexicon：a dash of sentiment beneath the surface meaning[C]. Sofia：The 51st Annual Meeting of the Association for Computational Linguistics，2013.
[32] Andreevskaia A，Bergler S. When specialists and generalists work Together：overcoming domain dependence in sentiment tagging[C]. Columbus：The 46th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies，2008.
[33] Lu Y E，Castellanos M，Dayal U，et al. Automatic construction of a context-aware sentiment lexicon：an optimization approach[C]. New York：The 20th International Conference on World Wide Web. Hyderabad India，2011.
[34] Xie R，Li C P. Lexicon construction：a topic model approach[C]. Yantai：The 2012 International Conference on Systems and Informatics，2012.
[35] 李慧. 面向学习体验文本的学习者情感分析模型研究[J]. 远程教育杂志，2021，39（1）：94-103.
[36] Zhang J F，Xia Y Q，Yao J M. A review towards microtext processing[J]. Journal of Chinese Information Processing，2012，4（26）：21-27，42.
[37] Yang A M，Zhou Y M，Lin J H. A method of Chinese texts sentiment classification based on Bayesian algorithm[J]. Applied Mechanics and Materials，2012，（263/266）：2185-2190.
[38] Ou X H，Cao Y，Mu X W. Classification of micro-blog sentiment based on naive Bayesian classifier[C]. Berlin：The 3rd International Conference on Logistics，Informatics and Service Science，2013.
[39] 林江豪，阳爱民，周咏梅，等. 一种基于朴素贝叶斯的微博情感分类[J]. 计算机工程与科学，2012，34（9）：160-165.
[40] Ren Y，Kaji N，Yoshinaga N，et al. Sentiment classification in resource-scarce languages by using label propagation[C]. Singapore：The 25th Pacific Asia Conference on Language，Information and Computation，2011.
[41] Escalante H J，Montes-y-Gómez M，Solorio T. A weighted profile intersection measure for profile-based authorship attribution[C]. Puebla：The 10th Mexican International Conference on Artificial Intelligence，2011.
[42] 孙广路，齐浩亮. 基于在线排序逻辑回归的垃圾邮件过滤[J]. 清华大学学报（自然科学版），2013，53（5）：734-741.
[43] 张莉，纪铭阳，胡宗玉，等. 基于随机森林和逻辑回归分类模型的烟叶精选品控指标筛选[J]. 江苏农业科学，2020，48（3）：214-217.
[44] 赵雅宏. 基于逻辑回归的供应商经营风险分析[D]. 济南：山东大学，2020.
[45] Jung J J. Maximum entropy-based named entity recognition method for multiple social networking services[J]. Journal of Internet Technology，2012，13（6）：931-937.
[46] 李佳儒，王玉珍，丁申宇. 基于逻辑回归的在线评论情感分类方法研究[J]. 东莞理工学院学报，2020，27（5）：50-54.
[47] 何跃，赵书朋，何黎. 基于情感知识和机器学习算法的组合微文情感倾向分类研究[J]. 情报杂志，2018，37（5）：189-194.

[48] Liu Z M，Liu L. Empirical study of sentiment classification for Chinese microblog based on machine learning[J]. Computer Engineering and Applications，2012，48（1）：1-4.
[49] Zhang L，Huang X Y，Jiang J，et al. CSLabel：an approach for labelling mobile app reviews[J]. Journal of Computer Science and Technology，2017，32（6）：1076-1089.
[50] Nimesh R，Veera Raghava P，Prince Mary S，et al. A survey on opinion mining and sentiment analysis[J]. IOP Conference Series：Materials Science and Engineering，2019，590（1）：012003.
[51] Ponte J M，Croft W B. A language modeling approach to information retrieval[J]. ACM SIGIR Forum，2017，51（2）：202-208.
[52] 刘艺梁，殷坤龙，刘斌. 逻辑回归和人工神经网络模型在滑坡灾害空间预测中的应用[J]. 水文地质工程地质，2010，37（5）：92-96.
[53] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.
[54] 薛福亮，刘丽芳. 基于 TF-IDF 和情感强度的细粒度情感分析：餐饮评论为例[J]. 信息系统工程，2020，（3）：83-84，86.

Research on the Construction of Negative Sentiment Dictionary in Finance Based on Financial News

ZHAO Youlin [1,2]，LIN Yini [2]，SUN Hong [2]，CHENG Lijie [2]，XU Jingnan [2]，LU Yingjun [3]
（1. School of Information Management，Nanjing University，Nanjing 210023，China；
2. School of Business，Hohai University，Nanjing 211100，China；
3. School of Information Management，Wuhan University，Wuhan 430072，China）

Abstract The emotional tendencies in the field of economics have their own domain-specific characteristics，and general sentiment dictionaries cannot accurately judge their emotional tendencies. This paper takes the real-time news information of the top 30 companies in the Dow Jones Industrial Average as the research object，sets up manual labeling rules，uses a variety of machine learning algorithms to train the labeled data sets，and uses the feature selection algorithm to extract the characteristic words as negative emotion words，so as to construct the negative emotion dictionary in the field of finance. Taking the current classic McDonald Financial Dictionary as the target sentiment dictionary，the accuracy of the domain dictionary constructed based on dynamic news information in this study is verified and evaluated by direct and indirect evaluation. 603 words have been successfully gotten in this research，and the results show that the coverage rate of the negative sentiment dictionary in the financial field constructed in this paper is 82.5%，and the accuracy of the recognition of financial news sentiment reaches 92.8%. The financial negative sentiment dictionary constructed in this paper can significantly improve the accuracy of negative sentiment tendency in finance.

Key words Negative emotion dictionary，Financial news，Characteristic words，Direct evaluation，Indirect evaluation

作者简介

赵又霖（1986—），女，河海大学商学院副教授、博士生导师，南京大学信息管理学院博士后，研究方向为空间数据分析与挖掘、知识组织。E-mail：sobzyl@hhu.edu.cn。

林怡妮（1999—），女，河海大学硕士研究生，研究方向为空间数据分析与挖掘。
孙虹（1996—），女，河海大学硕士研究生，研究方向为知识组织。
程丽洁（1996—），女，河海大学硕士研究生，研究方向为情感分析。
徐竟楠（1997—），女，河海大学硕士研究生，研究方向为知识组织。
陆颖隽（1963—），男，武汉大学信息管理学院副教授，研究方向为数字图书馆。

附录 基于财经新闻的金融领域负面情感词典结果

accuse	apart	adverse	alarm	alert	attack	anxiety	afraid	activist	argument
acute	argue	asymptomatic	beat	damage	barrel	bores	block	beat	busy
urge	black	complex	buffet	cancel	cancellation	disease	break	breakdown	bolson
blue	blood	bump	caution	complaint	contagious	communication	burden	closing	cancer
brutality	claim	challenge	charge	infectious	controversial	collapse	illness	competitive	cold
chuck	compensation	complicate	infect	launch	COVID	COVID-19	contraction	continent	combat
crash	crucial	dark	compromise	criminal	criticism	criticize	coronavirus	clinical	complication
crowd	crime	delay	cut	deadly	curb	curfew	death	dangerous	disaster
debate	crisis	deny	deadline	elderly	drag	down	decline	debt	downgrade
decrease	defense	devastate	disruption	difficult	difficulty	default	reduce	defeat	crude
deficit	din	dip	die	draft	division	discharge	exception	enforce	exit
epidemic	emergency	enforcement	disclose	drop	empty	dispute	extreme	disrupt	excess
evacuate	enemy	escalate	downside	emerge	drug	downward	downturn	halt	explosive
fell	expose	few	fail	fight	complications	filing	fraud	end	extremely
ford	fall	force	headline	hospitalization	flu	harm	downplay	delta	dis
gilead	grim	freeze	hike	hygiene	fatal	gap	grapple	hedge	violation
hurt	gamble	high-risk	historic	louisiana	gear	mass	hurricane	helicopter	icu
incident	hard	host	heavy	historical	greece	guideline	immigration	honor	hate
injury	hammer	identify	ignore	hotspot	michael	hardware	impossible	vaccination	victim
jersey	hardy	implement	implementation	illegal	heavily	hit	initially	immune	hybrid
interview	headwind	chase	independent	combine	kemp	jobless	intervention	metric	immunity
isolate	mandatory	inflation	influence	individual	historically	los	midst	competitor	infection
manufacturing	inflammatory	lack	lag	lam	infrastructure	interaction	layoff	late	loser
massachusetts	inmate	lending	lawmaker	lawsuit	intensive	investigation	lender	lead	kick
merck	merkel	lebanon	testify	legal	flow	lunar	madrid	hainan	league
midday	joseph	lockdown	long-term	loom	jump	larry	pacific	loss	leverage
mortality	kill	lung	luxury	macau	lamont	premium	maryland	manufacture	literally
paul	laura	morrison	management	manager	lawyer	limited	mockingly	mixed	matter
mitigate	oklahoma	minneapolis	migrant	mild	mandate	mayor	fake	missouri	misinformation
overwhelm	massive	mitigation	minnesota	minority	mark	weak	mortgage	narrow	michigan
payment	merger	multiple	moody	morgan	medium	military	numerous	mutual	mistake
portion	monetary	outline	outperform	operating	munching	musk	pandemic	outbreak	racism
murray	offset	perspective	participant	participate	notify	orlando	pneumonia	pending	overweight
prompt	option	plasma	panel	procter	operation	outstanding	poor	poise	panic
pursue	outlet	politician	pledge	pharmacy	originate	pakistan	potential	port	precaution
rally	pace	position	politics	plenty	plunge	penalty	pregnant	pound	prevent
regulatory	permit	pricing	prepared	presence	bigotry	pool	pressure	priority	portfolio
pilot	political	proceed	primary	rural	phase	publish	prolong	producer	powell

续表

representative	powerful	property	robust	profitability	untrusted	press	publish	protester	prevention
resurgence	procedure	range	quarantine	quarterly	presidential	prohibit	reasonable	rating	promote
reverse	professional	reaction	rank	rapid	refuse	protective	rock	rebound	protocol
rob	rival	recognize	reading	reality	processor	publicly	regret	reference	purpose
sample	push	regard	recommend	recommendation	profitable	race	roughly	regulation	rail
scrutiny	ramp	reject	symptoms	region	protection	rare	renew	relief	solid
separate	recession	repeat	repeatedly	remote	question	reduction	respect	reporting	regulator
sick	remark	retailer	retaliation	restart	remotely	reportedly	risky	reveal	represent
spokesperson	resign	rule	secretary	restructuring	resort	restriction	skinny	spike	resume
stake	responsibility	sander	senate	rush	restrict	risk	trajectory	liquidity	sweep
steve	retail	seattle	slide	saudi	retreat	ross	sensitive	segment	respiratory
strategy	review	selloff	sharply	shock	self-isolate	sad	severe	severely	scramble
switch	seasonal	shortly	slow	tension	sept	slip	slack	slam	severity
tariff	streak	simple	spark	shot	jerome	stretch	stadium	soar	shoot
treatment	slight	stick	underlie	statewide	single-day	suspend	tank	strike	strategist
violate	submit	stream	suspect	stockpile	slowly	syndrome	temporary	sudden	structure
resistance	stall	survive	transmit	subsequent	stabilize	storm	tout	tap	surround
against	symptom	throw	viral	tighten	surge	tally	unemployment	treasury	track
unclear	terrible	transmission	weaken	transparency	tackle	warning	vaccine	unable	tumble
blow	threshold	trend	volatility	trigger	temperature	tough	warm	unfortunately	trouble
underscore	bidden	unknown	widespread	unrest	timeline	transport	worried	victory	uncertain
aggressive	surgery	violence	virus	versus	transparent	typically	allegation	warn	weapon
weakness	withdraw	undermine							