# 金融市场收益率方向预测模型研究*

## ——基于文本大数据方法

顾文涛 王 儒 郑肃豪 杨永伟

**内容提要:** 金融市场的发展关系着一国的经济命脉,而股票市场作为金融市场的重要组成部分,对其收益率的研究也一直都是学术界的热点。财经新闻常被认为蕴含着丰富的信息,其中所包含的情感信息作为影响投资者投资决策的重要因素之一,对股票收益率也具有一定的影响。故本文构建了适用于金融投资领域的财经新闻情感词典来对财经新闻进行文本分析,同时构造了新的预测模型:将财经新闻文本中所含的情感量化为情绪指数并与时变密度函数相结合,得到时变加权密度模型。并在此基础上以模型评分为权重组合多个预测模型构建出评分加权模型用于股票收益率预测。结果显示,加入情绪指数能有效提高模型预测能力,而评分加权模型的预测能力则在此基础上更进一步,在准确率以及评分规则上基本达到双重最优。

**关键词:** 方向预测; 情绪指数; 评分加权

**DOI:** 10. 19343/j.cnki.11-1302/c.2020. 11. 006

**中图分类号:** C812 **文献标识码:** A **文章编号:** 1002-4565( 2020) 11-0068-12

# Research on The Prediction Model of The Direction of Financial Market Returns: Based on Text Big Data Method

Gu Wentao Wang Ru Zheng Suhao Yang Yongwei

**Abstract:** The development of the financial market concerns the economic lifeline of a country. As the stock market is an important part of the financial market, the research on the stock returns has always been a hot topic academically. Financial news is often considered to contain rich information, and as one of the important factors affecting investors' investment decisions, the emotion contained in financial news also has a certain impact on stock returns. Therefore, this paper constructs a sentiment lexicon of financial news, which is applicable in the financial investment field, and uses it to analyze the text of financial news. A new prediction model is established: to quantify the sentiment contained in financial news text as a sentiment index, and combine it with the time-varying density function to obtain a time-varying weighted density model. On this basis, scoring the model and using the model score as the weight to combine different prediction models to construct a scoring weighted model for stock returns prediction. The results show that the sentiment index of financial news can improve the prediction of the model, and the prediction effect of the scoring weighted model is double optimal in accuracy and scoring rules.

**Key words:** Direction Prediction; Sentiment Index; Scoring Weighted

* 基金项目: 浙江省一流学科 A 类项目"非参数密度函数在金融发展中的应用"( 1020JYN6517002G-01)。

# 一、引言

随着互联网的普及,文本数据海量增长,越来越多的学者开始采用文本大数据方法对金融市场进行研究。股票市场作为金融市场的重要组成部分,对其收益率的研究也一直都是学界的热点。在我国,股票市场的投资者以散户为主,这些投资者倾向于从财经新闻门户网站中获取投资信息。财经新闻作为投资信息的载体,多以文本数据的形式存在,其中蕴含的情感态度在很大程度上会对投资者的投资倾向产生影响,进而对股价的波动产生联动效应。因此对财经新闻文本信息进行分析,挖掘其中所包含的情感与情绪,对于股票市场收益率研究具有重要意义。

传统的文本分析方法多使用词典法。早在 20 世纪 60 年代初,Stone 和 Hunt( 1963) 就通过“the General Inquirer”语料库构建的哈佛词典进行文本分析方面的研究,该词典中包含了 3500 条用于表达语义的词条,并用正极性和负极性来表示情感极性。此后国外文本分析领域又陆续发布了一些情感词典,如: Wilson 等( 2005) 构建的 MPQA 主观情感词典和 Baccianella 等( 2010) 构造的 SentiWordNet 词典。相比而言,国内对于情感词典的研究起步较晚。HowNet 于 2007 年发布了“中英文情感分析用词语集( BETA) ”,共包含负面评价词语、正面评价词语、负面情感词语、正面情感词语、程度词语和主张词语六个大类。在这之后许多机构组织也构建了情感词典,如大连理工大学的情感词汇本体库,以及清华大学的清华褒贬义词典等。

上述提及的情感词典是目前进行文本分析时最常使用的通用型词典,然而这些词典的构建大多基于文学作品、媒体报道等,在特定领域的适用性和准确性存在一定的问题。例如,“负债”一词在金融领域不存在明显的情感倾向,但在一般语境下则含有负面情感。同时也有研究者发现通用型词典在特定领域的应用并没有达到预期的良好效果,如 Loughran 和 McDonald( 2011) 通过研究分析认为哈佛词典中 75%的消极词在金融领域的语义是非消极的。因此,当利用词典法对金融市场进行文本分析时,需要构建适用于金融领域的情感词典。

此外,股票的涨跌是投资的核心,因而对股票收益率方向的预测也就成为了人们关注的焦点。Gencay( 1999) ; Stock 和 Lambert( 2001) ; Bao 和 Yang( 2008) 都曾使用如 AR、ARMA 和多变量回归分析的统计方法来预测股票收益。Nyberg( 2011) 则是采用了二元动态 Probit 模型来预测美国月度股票收益率的方向。然而传统的参数预测模型大都需要较强的假设条件,这在一定程度上形成一种限制即如果真实分布与假设条件不符,模型的预测效果就会减弱,从而导致预测结果产生偏差。

目前,已有不少学者基于媒体新闻文本信息的分析来对金融市场进行研究。汪昌云和武佳薇( 2015) 采用了多家主流财经媒体的文本数据,运用词典法构建了媒体正、负面语气指数,并发现相比于正面媒体语气,负面媒体语气能够更好地解释 IPO 抑价率的变化。于琴等( 2017) 运用文本挖掘技术构建了新闻情绪指标,分析了研究新闻情绪与股票周收益率之间的关系。钟腾( 2018) 选取了我国七家具有广泛影响力的财经媒体,通过分析这些媒体的新闻文本数据,构建了媒体悲观指数。研究发现媒体报道情绪会显著影响投资者对于盈余信息的解读,即媒体悲观度越高,公告异常收益越低。不难看出,情绪指数是量化文本信息的关键,大多学者研究了情绪指数与金融市场收益的关系,而少有将其用于股票收益率方向的预测。

综上所述,本文主要有以下两点创新: 其一,构建了适用于金融投资领域的情感词典。以往利用词典法对金融文本进行的研究,由于未使用针对金融领域的专业性情感词典,容易导致分析结果与实际情感倾向之间有所出入。因此本文在权威词典的基础上,结合新闻文本构建了针对金融领

域的财经新闻情感词典，以提高文本分析的准确度。其二，将新闻文本的大数据分析和非参数模型相结合，构建了包含投资者情绪指数的非参数时变加权密度模型，提升了传统非参数模型的预测能力，丰富了大数据文本信息挖掘的应用研究。以往对于股票收益率方向的预测，以线性模型、ARMA 以及二元选择模型为主，这些模型的效果对假设条件的依赖性较大。故本文参考了 Gu 等(2018)的方法，改进并构建了包含投资者情绪指数的时变加权密度模型，将财经新闻文本中所包含的情感量化为情绪指数，作为影响因子加入到衰减因子的计算之中。这样有效地改善了原模型中新闻情感影响因子缺失的问题，使模型能够更好地预测股票收益率方向。除此之外，本文采用了准确率和概率评分规则作为预测模型性能的评价准则，然而在这种情况下会存在多个评价体系下同种模型产生不同结论这一问题。因此本文采用基于评分规则组合多种不同预测模型的方法，以此构建一种新的股票收益率方向的预测模型，并期望该模型在准确率和评分上都能有较好的表现。根据上述内容，本文结构可做如下安排：第二部分对投资者情绪指数进行构建，介绍时变非参数预测模型理论，同时构造一个新的模型——评分加权模型；第三部分以上证指数为股票收益率序列进行实证分析；第四部分对预测模型进行稳健性检验；第五部分进行总结。

## 二、新闻影响股票收益方向理论模型

### (一)财经新闻文本分析

财经新闻中蕴含着大量关于股市动向的资讯，投资者往往以这些资讯作为依据进行交易，而这些交易行为对股市的影响又以财经新闻的形式反馈给投资者。若将财经新闻资讯中所蕴含的情感倾向分为积极和消极，将股票市场的动向分为上涨和下跌，则可以认为在信息的不断交互后，股票市场的涨跌和财经新闻情感倾向将趋于一致。我国股市投资者主要由中小型投资者构成，他们缺乏一定的投资理论，且投机性强，极易受新闻情感倾向的影响。当财经新闻表现出积极的情感倾向时，投资者对形势抱以较为乐观的态度，大量买入从而造成股价上涨；而当新闻表现出消极的情感倾向时，投资者又急于抛售风险从而导致股票价格下跌。结合相关研究，图 1 可用来对财经新闻与股市的关系进行描述。

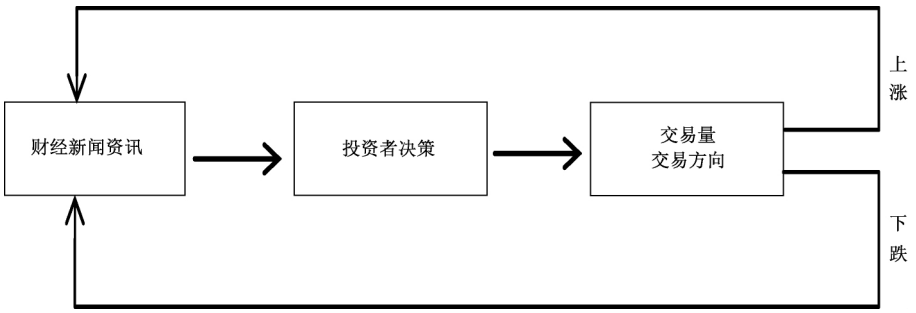


图 1 新闻与股市关系图

随着互联网的迅猛发展，目前财经新闻多以文本数据的形式存在，这些数据中有价值的信息无法直接被计算机读取、计算，因此要进行文本处理后才能挖掘其价值。本文主要考虑的是财经新闻中所蕴含的情感信息，每篇新闻文本都是由多个词语组成的信息集合，其中某些词语含有一定的情感倾向(称为情感词)，则一篇新闻文本的情感可由情感词来体现。

情感词典法是一种常用的文本分析方法，该法通过识别文本中的情感词，并对这些词进行计算进而判别文本的情感倾向。知网的 HowNet 词典、大连理工大学的情感词汇本体库、清华大学的清华褒贬义词典，以及台湾大学 NTUSD 简体中文情感极性词典等都是常用的中文通用型情感词典。

然而,通用型词典并不适用于所有情况,尤其在专业性比较高的领域,通用词典中部分词的情感标注并不适用。例如,"庄"字在一般语境中并不具备特殊的情感倾向,但在股票方面则带有强烈的负面情绪。针对这个问题,许多学者构建了适用于特定领域的中文词典。汪昌云和武佳薇(2015)基于《现代汉语词典》、《最新汉英经济金融常用术语使用手册》、Loughran-McDonald 词典[①]中文版和知网-中文信息结构库,构建了一个适用于我国财经新闻媒体的正负面词库。彭红枫和林川(2018)基于 Loughran-McDonald 词典中文翻译版,通过人工筛选并结合中文版 LIWC 词典构建了 P2P 网络借贷词典。这些已有的特定领域词典多以人工筛选、词典重组为主要方式进行构筑,但是人工筛选较为费时费力,且受主观性影响较大,故本文考虑使用软件筛选和人工筛选相结合的方式进行情感词典的构建。目前 Word2Vec 在计算语言学等领域得到了广泛的应用,但在金融领域的应用相对较少。它实际上是一个简化的神经网络,将文本词汇投射到一个向量空间中,并赋予每个词一个词向量,通过计算向量余弦相似度来衡量词语间的相似性。因此,本文以 HowNet 词典以及翻译后的 Loughran-McDonald 词典为基础,同时利用 Word2Vec 和人工筛选的方式对词典进行扩充。词典构建方法基本思路如图 2 所示:

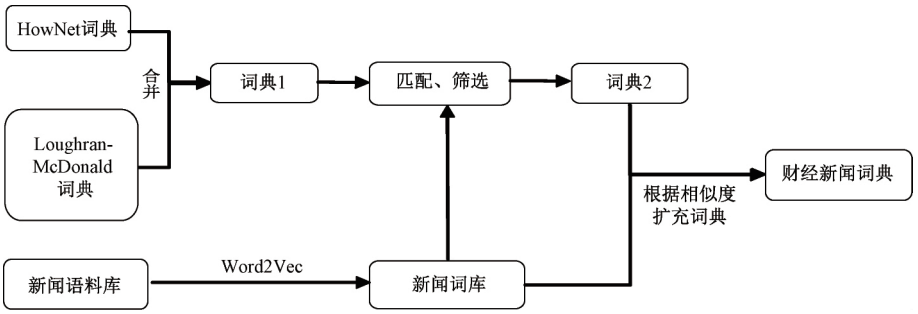


**图 2　财经新闻情感词典(FNSD)构建思路图**

①HowNet 词典包含了正负面情感词以及正负面评价词,因为字相对于词语而言其情感传递能力较弱,故去除词典中字符数为 1 的单个字。

②目前在金融领域具有权威性的情感词典为 Loughran-McDonald 词典,但由于英文与中文表达存在差异,直接翻译而来的词典并不适用于中文文本。故根据财经英汉词典,将 Loughran-McDonald 词典翻译成中文,若某个词语出现在 Loughran-McDonald 词典中,但财经英汉词典并未收录该词,则可利用在线翻译平台对该词进行翻译。

③将步骤①和步骤②中处理过的词典合并去重,形成词典 1。

④将从新浪财经上爬取的 2016 年度新闻文本作为语料库,先对语料库中语料进行分词、去停用词和去符号等预处理操作,然后利用 Word2Vec 对语料库进行训练,得到词向量空间。同时统计所有词的词频,设定低频阈值为 20,去除词频小于 20 的词汇,得到新闻词库。此后将该词库与词典 1 进行匹配,之后再对剩余词语进行人工筛选,剔除不适用财经金融领域的词语,形成词典 2。

⑤通过计算词向量余弦相似度,基于新闻词库对词典 2 进行扩充(本文设定相似度水平需大于 80%),再经过去重等操作,最终得到针对金融领域的财经新闻情感词典。

通过上述方法构建的财经新闻情感词典将作为本文所使用的词典,它与 HowNet 词典相比,包含了更多财经以及金融领域常用的词汇。

词典完成之后需对文本中的情感进行量化。Antweiler 和 Frank(2004)提出了"bullishness"情

① Loughran-McDonald 词典是由 Loughran 和 McDonald(2011)构建的适用于金融领域的英文词典。

绪指数用以量化文本情绪，并发现情绪与同期股票收益率显著正相关。此后许多学者在研究情绪与股票收益相关问题时均参照该法构建情绪指数，如陈浪南和苏湃(2017)；梅立兴等(2019)。故本文也采用该方式构建投资者情绪指数(Sentiment Index)：

$$SI = \log \frac{1 + N^p}{1 + N^n} \tag{1}$$

其中，$N^p$ 表示新闻文本中积极词数，$N^n$ 表示新闻文本中消极词数。由式(1)可知，当积极词汇相对于消极词汇越多时，情绪指数越高，表明投资者情绪越高涨，而当消极词汇相对较多时，则表明投资者情绪较为低落。需要注意的是，本文所分析的语料并非单个新闻文本，而是以日为单位定义新闻语料库，基于单日内所有语料文本构建情绪指数。这是因为不同新闻文本存在篇幅不平衡的情况，通过这样的方式能够有效避免单篇新闻文本情绪指数极高或极低。

(二)股票收益率预测模型理论框架

传统的股票收益率预测模型多为线性模型，而这些模型都存在潜在的假设，如方向预测问题，经典的 Logistic 模型就对方向做出了服从 Logistic 分布的假设。但股票市场受众多因素影响，具有极高的复杂性和不确定性，数据实际的分布往往与预先假定的分布存在着较大区别。为避免模型预设，本文使用非参数的方法以投资者情绪指数为影响因子构建了非参数时变加权密度模型，同时基于评分规则组合多个预测模型构造了评分加权模型，并和 Logistic 模型进行了对比。下面将按照时变密度函数模型(TVF)、时变加权密度模型(F-TVF)、模型评价以及评分加权模型的顺序进行介绍。

1.时变密度函数模型(TVF)。

经典核密度估计会为每期观测值分配相等的权值，而 Harvey 和 Oryshchenko(2012)在核密度估计的基础上引入指数加权移动平均法，考虑赋予离当期值越近的观测值一个更高的权重，从而构建了时变概率密度函数。假设 $y_1,\cdots,y_T$ 为一时间序列观测值，则指数加权移动平均法的表示如下：

$$m_t = \sum_{i=0}^{t-1} y_{t-i} w_{t,i}, t = 1,\cdots,T \tag{2}$$

其中 $w_{t,i} = (1-\omega)\omega^i$ 为权重($0 \leqslant \omega < 1$)。$\omega$ 为衰减因子，表示对不同期观测值赋予的权重的衰减速度。同时式(2)可以写成如下递归形式：

$$m_{t+1|t} = m_{t|t-1} + (1-\omega) v_t \tag{3}$$

且 $v_t = y_t - \hat{y}_{t|t-1}$，表示向前一步预测误差。

将该法与核密度估计相结合，有：

$$\hat{f}_t(y) = \frac{1}{h}\sum_{i=1}^{t} K\left(\frac{y - y_i}{h}\right) w_{t,i}, t = 1,\cdots,T \tag{4}$$

其累积分布函数为，

$$\hat{F}_t(y) = \sum_{i=1}^{t} H\left(\frac{y - y_i}{h}\right) w_{t,i}, t = 1,\cdots,T \tag{5}$$

其中 $K(\cdot)$ 为核函数，$h$ 为核函数的窗宽，$H(\cdot)$ 为与 $K(\cdot)$ 相对应的累积核函数。$w_{t,i}$ 为权重，且有 $\sum_{i=1}^{t} w_{t,i} = 1, w_{t,i} = (1-\omega)\omega^{t-i}$。与式(3)类似，累积分布函数也可写成如下递归形式：

$$\hat{F}_{t+1|t}(y) = \hat{F}_{t|t-1}(y) + (1-\omega) V_t(y), t = 1,\cdots,T \tag{6}$$

同样概率密度函数为：

$$\hat{f}_{t+1|t}(y) = \hat{f}_{t|t-1}(y) + (1-\omega)\, v_t(y),\ t = 1,\cdots,T \tag{7}$$

其中，$V_t(y) = H\left(\frac{y-y_t}{h}\right) - \hat{F}_{t|t-1}(y)$，$v_t(y) = \frac{1}{h}K\left(\frac{y-y_t}{h}\right) - \hat{f}_{t|t-1}(y)$。如上，式(6)与式(7)称为时变密度函数模型，简记为 TVF。

2.时变加权密度模型(F-TVF)。

与线性模型相比，非参数时变密度函数较好地解决了模型误设问题。但是在实际情况中，影响股票数据的因素多种多样，所以 Gu 等(2018)考虑在密度估计之中加入影响因素。

设 $X$ 是与 $Y$ 相关的影响因子，在模型权重部分加入该因子，则衰减因子可表示为:

$$\tilde{\omega} = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \tag{8}$$

当 $X$ 对 $Y$ 影响较大时，能够得到较大的 $\tilde{\omega}$，且满足 $0 \leqslant \tilde{\omega} < 1$。那么 $Y$ 的密度估计则可表示为:

$$\hat{f}_t(y) = \frac{1}{h}\sum_{i=1}^{t} K\left(\frac{y-y_i}{h}\right)\tilde{w}_{t,i},\ t = 1,\cdots,T,\ \tilde{w}_{t,i} = (1-\tilde{\omega})\,\tilde{\omega}^{t-i} \tag{9}$$

该式同样可以写成递归形式，

$$\hat{f}_{t+1|t}(y) = \hat{f}_{t|t-1}(y) + (1-\tilde{\omega})\, v_t(y),\ t = 1,\cdots,T \tag{10}$$

上式(10)即为非参数时变加权密度模型，简记为 F-TVF。考虑到新闻情绪对股票的影响，本文将情绪指数作为影响因子加入到衰减因子的计算之中，对原有模型进行了改进，使模型能够更好地预测股票收益率方向。

3.模型预测性能评价。

(1)收益方向设定。若 $y_t$ 表示 $t$ 期的股票收益率，则收益方向为: 当 $y_t > 0$ 时，真实收益率为上涨，记为 $I_t = 1$; 当 $y_t \leqslant 0$，真实收益率为下跌，记为 $I_t = -1$。

根据前文所介绍的 TVF 模型和 F-TVF 模型可以得出下一期股票收益率上涨的概率公式为:

$$P = Prob(y_{t+1} > 0) = 1 - \hat{F}_{t+1|t}(0) \tag{11}$$

相似地，股票收益率下跌的概率为:

$$P^d = Prob(y_{t+1} \leqslant 0) = \hat{F}_{t+1|t}(0) \tag{12}$$

其中，上标 $d$ 表示股票收益率下跌，且 $0 \leqslant P, P^d < 1, P + P^d = 1$。设定阈值 $L$，$P_i$ 表示在 $T+i$ 期股票收益上涨的概率，那么股票收益方向的预测值则可以写成如下形式:

$$I_i^f = \begin{cases} 1, & if\ P_i > L \\ -1, & if\ P_i \leqslant L \end{cases},\ i = 1,\cdots,N \tag{13}$$

同理，根据真实股票收益率的涨跌，有:

$$I_i = \begin{cases} 1, & if\ y_{T+i} > 0 \\ -1, & if\ y_{T+i} \leqslant 0 \end{cases},\ i = 1,\cdots,N \tag{14}$$

其中，$I_i^f$ 为预测的收益率方向，$I_i$ 为真实的收益率方向。

(2)模型评价。准确率 CR 常被用作方向预测的评判标准，其表达式如下:

$$CR = \frac{I^{uu} + I^{dd}}{I^{uu} + I^{ud} + I^{dd} + I^{du}} \tag{15}$$

其中，

$$I^{uu}=\sum_{i=1}^{N}1(I_i^f=1,I_i=1),I^{ud}=\sum_{i=1}^{N}1(I_i^f=1,I_i=-1)$$

$$I^{du}=\sum_{i=1}^{N}1(I_i^f=-1,I_i=1),I^{dd}=\sum_{i=1}^{N}1(I_i^f=-1,I_i=-1)$$

但是对于方向预测，将准确率作为唯一的评价标准具有一定的缺陷。考虑如下情况，模型 I 和模型 II 同时用于股票收益率的预测，其向前一步预测的结果如表 1 所示，其中 $I$ 表示收益率真实方向，$P_I$ 为模型 I 预测上涨的概率，$P_{II}$ 为模型 II 预测上涨的概率。

表 1 模型 I、II 预测结果

| | 第一期 | 第二期 | 第三期 | 第四期 | 第五期 |
|---|---|---|---|---|---|
| $I$ | 1 | 1 | −1 | −1 | 1 |
| $P_I$ | 0.83 | 0.81 | 0.13 | 0.82 | 0.74 |
| $P_{II}$ | 0.56 | 0.52 | 0.58 | 0.44 | 0.68 |

若将阈值 $L$ 设定为 0.5，可以计算得到模型 I 和模型 II 的准确率均为 0.8；若将阈值设定为 0.4，可以发现模型 I 的准确率并未发生改变，而模型 II 的准确率由 0.8 变为了 0.6，此时判定模型 I 的预测性能要优于模型 II。由此可见阈值 $L$ 大小的设定对准确率影响很大，同个模型的准确率在不同阈值下发生了改变，这使得模型的评价具有极大的不确定性。探其究竟，这主要是由于准确率没有包含概率的信息，即期望的模型概率应符合如下情况：如果模型预测上涨，那么这个上涨概率最好尽可能趋近于 1，如果预测下跌，则希望预测上涨的概率尽可能趋向于 0。

因此本文引入概率评分规则，作为除准确率之外用于评价模型优劣的准则，相比于准确率来说，概率评分规则还考虑了概率因素。本文采用了常见的二次评分和对数评分规则。

基于向前一步预测，根据预测模型可得概率 $(P,P^d)$，其中 $P$ 为上涨的概率，$P^d$ 为下跌的概率。则对数评分 $S^l$ 可表示为：

$$S^l=\begin{cases}\log(P), & I=1\\ \log(P^d), & I=-1\end{cases} \tag{16}$$

同时本文参考 Pauwels 和 Vasnev(2017) 的构造方法得到的二次评分规则可表示为：

$$S^q=\begin{cases}2P-P^2-(P^d)^2, & I=1\\ 2P^d-(P^d)^2-P^2, & I=-1\end{cases} \tag{17}$$

由式(16)和式(17)可得，在上述两种评分规则下均有一致的模型评判标准，即概率评分值越高表示模型预测效果越好。

若预测样本时间段为 $(\tau_1,\tau_2]$，每一个预测点都可计算一个评分，则 $(\tau_1,\tau_2]$ 期间的平均二次评分和平均对数评分如下所示：

$$\bar{S}^q=\frac{1}{\tau_2-\tau_1}\sum_{t=\tau_1+1}^{\tau_2}S_t^q,\bar{S}^l=\frac{1}{\tau_2-\tau_1}\sum_{t=\tau_1+1}^{\tau_2}S_t^l \tag{18}$$

其中 $S_t^q$ 表示按照二次评分规则得到的 $t$ 期的评分值，$S_t^l$ 表示按照对数评分规则得到的 $t$ 期的评分值。

4.评分加权模型。

由于不同的评价方式对同一个预测模型可能存在着不同的判定结果，因此本文采用以概率评分值为权重对各个模型进行加权的方式构建一个新的预测模型，以期新模型在准确率和评分方面均具有良好的表现。

根据训练期内各模型的不同评分，基于二次评分和对数评分规则为模型分别构造组合权重

$$w_i^{q} = \frac{\bar{S}_i^{q}}{\sum_{i=1}^{M} \bar{S}_i^{q}}, w_i^{l} = \frac{1/|\bar{S}_i^{l}|}{\sum_{i=1}^{M} 1/|\bar{S}_i^{l}|}, i = 1, \cdots\cdots, M \tag{19}$$

其中 $M$ 表示共有 $M$ 个预测模型，且 $\sum_{i=1}^{M} w_i = 1$，$\bar{S}_i^{q}$ 表示预测模型 $i$ 的平均二次评分，$\bar{S}_i^{l}$ 表示预测模型 $i$ 的平均对数评分。由于对数评分规则返回为负值，因而将权重设置为对数评分的绝对值并取导数。假设训练期时段为 $(\tau_1, \tau_2]$，那么 $\tau_2 + 1$ 期可由训练期的平均分值进行预测，其预测概率密度估计如下式:

$$\hat{f}_{t+1|t}(y) = \sum_{i=1}^{M} \hat{f}_{t+1|t,i} w_{t,i} \tag{20}$$

其累积分布形式为:

$$\hat{F}_{t+1|t}(y) = \sum_{i=1}^{M} \hat{F}_{t+1|t,i} w_{t,i} \tag{21}$$

且基于评分的加权模型预测上涨的概率可写为:

$$\begin{aligned} P_{t+1} &= 1 - \hat{F}_{t+1|t}(0) = 1 - \sum_{i=1}^{M} \hat{F}_{t+1|t,i}(0)\, w_{t,i} \\ &= \sum_{i=1}^{M} (1 - \hat{F}_{t+1|t,i}(0))\, w_{t,i} = \sum_{i=1}^{M} P_{t+1,i} w_{t,i} \end{aligned} \tag{22}$$

本文将基于两种评分规则所构建的评分加权模型分别称为二次评分加权模型(Quad)和对数评分加权模型(Log)。

综述预测模型理论部分可知，TVF 模型考虑赋予离当期值越近的观测值一个更高的权重，以反映不同期观测值对预测的影响; F-TVF 模型则是在 TVF 模型的基础上考虑在密度估计中加入影响因素来提升模型的预测精度; 而评分加权模型的内在逻辑是对训练期内评分值高的模型予以更高的权重，通过模型组合加权的方式提高模型的预测性能。

## 三、股票收益率方向预测实证结果分析

### (一) 投资者情绪指数可预测性分析

新浪财经基于全方位服务各类投资者这一理念，与各路领先财经新闻媒体通力合作，目前已成为国内知名度较高且影响力较大的专业财经平台，是众多投资者获取最新市场资讯的不二选择。因此本文采用 Python 爬虫爬取了该网站“股市及时雨”和“市场研究”两个板块 2016 年至 2018 年 1 月的新闻文本，共计 5 万余篇新闻。同时以日为单位定义新闻语料库，基于单日内所有语料文本构建情绪指数。

本文以上证指数作为研究对象，获取 2017 年 1 月 3 日至 2018 年 1 月 31 日共计 271 个交易日的样本。采用公式 $y_t = 100 \times \ln(P_t/P_{t-1})$ 计算日收益率序列，其中 $y_t$ 为日收益率，$P_t$ 为 $t$ 期的收盘价。同时采用滚动时间窗口来预测衡量模型样本外的预测情况，将数据分为测试集与训练集，其中测试集样本量为 60。

对上证指数收益率进行正态性检验以及描述性统计分析，结果如表 2 所示。由表 2 可得 S-W(Shapiro-Wilk)和 K-S(Kolmogorov-Smirnov)检验的 P 值均小于 0.05，说明上证指数收益率并不服从正态分布，因此采用非参数的方法能够更好地捕捉这种非正态性时间序列的特性。此外对上证指数收益率序列进行 ADF 单位根检验，结果显示序列平稳。

为检验文本构建的情感词典与情绪指数的有效性，建立如下回归模型:

表 2 上证指数收益率正态性检验及描述性统计结果

| S-W | K-S | ADF | Mean | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.01 | 0.04 | 0.31 | -0.29 | 1.24 |

注: 表中前三列所对应的值均为检验的 P 值

$$y_t = \alpha + \sum_{i=0}^{n_1} \beta_i SI_{t-i} + \sum_{i=1}^{n_1} \gamma_i y_{t-i} \quad (23)$$

其中，$y_t$ 为上证指数收益率，$SI_{t-i}$ 表示滞后 $i$ 期的情绪指数。对情绪指数进行单位根检验，结果显示序列一阶差分平稳。表 3 为基于不同词典构建的情绪指数与上证指数收益率的回归结果。由表 3 可知，只有财经新闻情感词典的情绪指数在 1%水平上显著，而其他通用型词典的回归系数均不显著。这说明了本文构建词典的有效性，且与通用型词典相比更适用于金融领域的分析。同时由表中第一行结果可知，情绪指数与收益率存在正相关性。

表 3 基于不同词典构建的情绪指数对股票收益预测回归结果

| | *Cons* | $y_{t-1}$ | $y_{t-2}$ | $SI_t$ | $SI_{t-1}$ | $SI_{t-2}$ |
|---|---|---|---|---|---|---|
| *FNSD* | 0.039 | -0.033 | 0.040 | 0.971* | 0.524*** | 0.309 |
| *HOWNET* | 0.034 | 0.003 | 0.011 | 1.302 | -0.609 | 0.032 |
| *TSINGHUA* | 0.034 | 0.018 | -0.001 | -0.193 | 0.002 | 0.045 |
| *DLUTSD* | 0.033 | 0.024 | 0.001 | -0.566 | -0.671 | 0.527 |

注: *FNSD* 表示本文所构建的财经新闻情感词典，*HOWNET* 表示知网词典，*TSINGHUA* 表示清华褒贬义词典，*DLUTSD* 表示大连理工情感词汇本体库; *** 表示在 10%水平下显著，* 表示在 1%水平下显著。

### (二) 基于预测模型的评价结果

本文采用四种预测模型对上证指数收益率序列的涨跌进行预测，分别为 Logistic 模型、TVF 模型、F-TVF 模型以及评分加权模型。其中评分加权模型又分为对数评分加权模型和二次评分加权模型，它们是根据 Logistic、TVF、F-TVF 这三种模型基于两种不同评分规则进行加权组合得到的。为避免阈值 $L$ 的设定对准确率的影响，本文设置了多个阈值，且以准确率和评分规则两种方式对每个模型进行评价。各模型预测结果如表 4 所示，同时为了更加直观地对比各模型的预测效果，表 5 给出了对应的预测效果排序。

表 4 不同模型样本外预测结果

| *L* | Quad | Log | F-TVF(SI) | TVF | Logistic |
|---|---|---|---|---|---|
| **各模型预测准确率** | | | | | |
| 0.400 | 0.63 | 0.63 | 0.57 | 0.57 | 0.58 |
| 0.425 | 0.64 | 0.63 | 0.58 | 0.56 | 0.56 |
| 0.450 | 0.61 | 0.62 | 0.57 | 0.54 | 0.56 |
| 0.475 | 0.57 | 0.58 | 0.57 | 0.56 | 0.55 |
| 0.500 | 0.57 | 0.58 | 0.55 | 0.54 | 0.53 |
| **各模型预测评分** | | | | | |
| LOG | -0.69 | -0.69 | -0.72 | -0.73 | -0.74 |
| QUAD | 0.50 | 0.50 | 0.49 | 0.47 | 0.46 |

注: F-TVF(SI) 即表示包含投资者情绪指数的非参数时变加权密度模型，LOG 表示对数评分规则，QUAD 表示二次评分规则

结合表 4 及表 5，对各个模型的预测结果从准确率层面进行分析，得到的结论如下:

第一，显然所有模型的预测准确率均大于 0.5，这说明相较于随机判断，本文构建的模型均具有一定的预测判断性能。同时不难发现，所有模型的准确率的最大值均出现在阈值 0.4 至 0.425 左右，且 Log 模型和 Quad 模型的预测准确率变化范围要略大于其他三个模型。

第二，基本在每一个阈值层面下，包含了投资者情绪指数的 F-TVF 模型的准确率均优于 TVF 模型，这表明构建情绪指数，即考虑新闻媒体情感倾向的影响，的确能有效提升模型的预测能力。

表 5　　不同模型样本外预测结果排序

| *L* | Quad | Log | F-TVF( SI) | TVF | Logistic |
|---|---|---|---|---|---|
| **各模型预测准确率** | | | | | |
| 0. 400 | 1 | 1 | 4 | 4 | 3 |
| 0. 425 | 1 | 2 | 3 | 4 | 4 |
| 0. 450 | 2 | 1 | 3 | 5 | 4 |
| 0. 475 | 2 | 1 | 2 | 4 | 5 |
| 0. 500 | 2 | 1 | 3 | 4 | 5 |
| **各模型预测评分** | | | | | |
| LOG | 1 | 1 | 3 | 4 | 5 |
| QUAD | 1 | 1 | 3 | 4 | 5 |

第三，虽然 Log 模型和 Quad 模型的预测准确率较为接近，但是相比于其他三种模型，这两个评分加权模型的预测准确率均得到了较大的提升。

对各个模型的预测结果从评分规则层面进行分析，可以看到无论是基于哪种评分规则，Logistic 模型的模型评分均为最低，而本文构建的评分加权模型的评分则达到最高。这表明在考虑概率因素的情况下，评分加权模型要优于其他三种预测模型，而包含情绪指数的 F-TVF 模型要优于不考虑投资者情绪影响的模型，即 Logistic 和 TVF 模型。

综合上述分析结果可得，考虑到新闻媒体情感倾向对投资者的影响而引入投资者情绪指数的确能有效提高模型的预测准确率。具有一定的实际意义。同时，本文构建的评分加权模型在准确率和概率评分方面均有较好的表现，这体现出了评分加权模型的优越性。

( 三) 交易策略模拟

本文尝试从经济角度对各模型的预测能力进行评估，基于预测模型对预测结果进行模拟交易。考虑到夏普比率( SR) 是基金绩效评价标准化指标，对投资者进行交易策略的选择具有一定的指导意义，故与年化收益率( Rc) 同作为评价标准。交易策略模拟结果如表 6 所示。

表 6　　交易策略模拟结果

| *L* | Quad | | Log | | F-TVF( SI) | | TVF | | Logistic | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rc( %) | SR | Rc( %) | SR | Rc( %) | SR | Rc( %) | SR | Rc( %) | SR |
| 0. 400 | 22. 95 | 0. 12 | 22. 95 | 0. 12 | 12. 53 | 0. 06 | 11. 71 | 0. 06 | 11. 23 | 0. 06 |
| 0. 425 | 25. 36 | 0. 13 | 23. 88 | 0. 12 | 14. 29 | 0. 07 | 8. 86 | 0. 05 | 3. 39 | 0. 02 |
| 0. 450 | 21. 90 | 0. 12 | 23. 96 | 0. 13 | 13. 53 | 0. 07 | 7. 50 | 0. 04 | 2. 06 | 0. 01 |
| 0. 475 | 17. 20 | 0. 09 | 18. 03 | 0. 10 | 12. 59 | 0. 07 | 11. 70 | 0. 07 | 2. 56 | 0. 01 |
| 0. 500 | 14. 50 | 0. 07 | 15. 15 | 0. 09 | 8. 44 | 0. 06 | 5. 84 | 0. 04 | -1. 17 | 0. 00 |

由上表可得，在同一阈值层面下，各模型的交易策略模拟表现出与模型预测能力评价相似的结果，即 TVF 模型的交易策略要优于 Logistic 模型，基于情绪指数的 F-TVF 模型要优于 TVF 模型，两个评分加权模型的交易策略效果相差不大，但都优于 F-TVF 模型。

## 四、稳健性检验

本文选取基金指数和 A 股指数对模型进行稳健性检验。基金指数通常能够反映金融市场的综合变化，通过对其预测分析可研究某个行业或市场受财经新闻的影响程度。财经新闻主要反映的是我国 A 股相关情况，A 股指数的变化与股市行情的变化几乎是同步的，它是投资者判断股票变化趋势的重要参考依据。由表 7 和表 8 可知，通过比较各模型间的准确率与模型评分可得出与上文较为相似的分析结果，说明本文构建的预测模型具有一定的稳健性。

表 7 基金指数预测结果

| *L* | Quad | Log | F-TVF( SI) | TVF | Logistic |
|---|---|---|---|---|---|
| **各模型预测准确率** | | | | | |
| 0. 400 | 0. 57 | 0. 56 | 0. 56 | 0. 55 | 0. 54 |
| 0. 425 | 0. 56 | 0. 56 | 0. 55 | 0. 55 | 0. 52 |
| 0. 450 | 0. 55 | 0. 55 | 0. 53 | 0. 53 | 0. 51 |
| 0. 475 | 0. 54 | 0. 54 | 0. 53 | 0. 51 | 0. 49 |
| 0. 500 | 0. 50 | 0. 50 | 0. 51 | 0. 50 | 0. 48 |
| **各模型预测评分** | | | | | |
| LOG | -0. 72 | -0. 72 | -0. 76 | -0. 77 | -0. 77 |
| QUAD | 0. 48 | 0. 48 | 0. 46 | 0. 44 | 0. 43 |

表 8 A 股指数预测结果

| *L* | Quad | Log | F-TVF( SI) | TVF | Logistic |
|---|---|---|---|---|---|
| **各模型预测准确率** | | | | | |
| 0. 400 | 0. 59 | 0. 59 | 0. 56 | 0. 54 | 0. 57 |
| 0. 425 | 0. 62 | 0. 61 | 0. 57 | 0. 55 | 0. 56 |
| 0. 450 | 0. 59 | 0. 59 | 0. 57 | 0. 54 | 0. 56 |
| 0. 475 | 0. 55 | 0. 56 | 0. 56 | 0. 55 | 0. 55 |
| 0. 500 | 0. 52 | 0. 52 | 0. 56 | 0. 53 | 0. 53 |
| **各模型预测评分** | | | | | |
| LOG | -0. 71 | -0. 71 | -0. 74 | -0. 74 | -0. 75 |
| QUAD | 0. 48 | 0. 48 | 0. 47 | 0. 46 | 0. 45 |

## 五、结论

本文以财经新闻文本作为研究对象，构建财经新闻情感词典，将财经新闻文本中的情感量化为情绪指数，构造了一个结合情绪指数的非参数预测模型来提高模型的预测能力。对于预测模型，本文采用了改进的非参数时变加权密度模型进行预测，同时以准确率和概率评分规则作为评价标准。为了达到评分与准确率的双重最优，还基于两种评分规则( 对数评分以及二次评分) 构建了评分加权模型，并且以上证指数作为实证样本进行了分析验证。根据上述方法和实证结果得到的结论有:

第一，本文构建的情感词典与通用型词典相比更适用于金融领域，且构建的投资者情绪指数与股票收益率呈现正相关性。这说明投资者易受到新闻情感倾向的影响，当新闻情绪较为高涨时，大量投资者为了抓住时机就会跟随自己的主观意识买入股票，从而造成股价上涨。

第二，本文将投资者情绪指数作为影响因子加入到预测模型中，实证分析后发现相较于 Logistic 和 TVF 模型，加入了情绪指数的 F-TVF 模型在预测准确率及评分方面都有了一定的提升，这说明考虑新闻媒体情感倾向的影响能有效提升模型的预测能力。

第三，本文引入概率评分规则，作为除准确率之外用于评价模型优劣的准则。同时为了使模型在准确率与评分上得到双重最优。本文以评分分值为权重对各个模型进行加权，实证结果表明基于两种评分规则构建的模型要优于 Logistic、TVF 以及 F-TVF 模型。

以往的文本信息研究往往需要人工识别、标注和分析文本数据，因此会耗费大量的人力和时间，且由于人工处理的主观性较强，没有既定且统一的标准，研究成果很难复刻。而随着文本挖掘技术的不断进步，许多机器学习、神经网络的方法被应用于文本分析当中，大大提升了分析的效率和效果。这些方法能够更为精准快速地挖掘出海量网络文本中的有效信息，将其用于财经新闻文本的分析则能够更为细致地刻画股市动向，使投资者能够做出更为合理的决策推断。同时，将新闻文本的大数据分析和非参数模型相结合，这提供了一种新的研究思路，并且丰富了大数据文本信息挖掘的应用研究。最后，本文通过对现有情感词典进行整合、筛选和扩充的方式构建了财经新闻情感词典，该法尚不完善，效果也有待提高。后续会综合考虑文本语料的其他特征，对财经新闻情感

词典的构建方法进行优化,以提升词典的针对性和精准度。

## 参考文献


[1] 陈浪南,苏湃. 社交媒体对股票市场影响的实证研究 [J]. 投资研究, 2017,36(11): 17-35.

[2] 梅立兴,张灿,何鲁. 投资者情绪与股票收益——来自移动互联网的实证研究 [J]. 南方经济, 2019(3): 36-53.

[3] 彭红枫,林川. 言之有物: 网络借贷中语言有用吗? ——来自人人贷借款描述的经验证据 [J]. 金融研究, 2018(11): 133-152.

[4] 汪昌云,武佳薇. 媒体语气、投资者情绪与 IPO 定价 [J]. 金融研究, 2015(9): 174-189.

[5] 于琴,张兵,虞文微. 新闻情绪是股票收益的幕后推手吗 [J]. 金融经济学研究, 2017,32(6): 95-103.

[6] 钟腾. 媒体报道情绪、投资者解读与股票市场稳定 [J]. 金融监管研究, 2018(5): 32-53.

[7] 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.

[8] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining [C]. LREC, 2010(10): 2200-2204.

[9] Bao D, Yang Z. Intelligent Stock Trading System by Turning Point Confirming and Probabilistic Reasoning [J]. Expert Systems with Applications, 2008,34(1): 620-627.

[10] Gencay R. Linear, Non-Linear and Essential Foreign Exchange Rate Prediction with Simple Technical Trading Rules [J]. Journal of International Economics, 1999,47: 91-107.

[11] Gu W, Yang Y, Liu Z. Forecasting Stock Returns Based on a Time-Varying Factor Weighted Density Model [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2018,22(6): 831-837.

[12] Harvey A, Oryshchenko V. Kernel Density Estimation for Time Series Data [J]. International Journal of Forecasting, 2012,28(1): 3-14.

[13] 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.

[14] Nyberg H. Forecasting the Direction of the US Stock Market with Dynamic Binary Probit Models [J]. International Journal of Forecasting, 2011,27(2): 561-578.

[15] Pauwels L L, Vasnev A L. Forecast Combination for Discrete Choice Models: Predicting FOMC Monetary Policy Decisions [J]. Empirical Economics, 2017,52(1): 229-254.

[16] Stock J R, Lambert D M. Strategic Logistics Management [M]. Published by McGraw-Hill Higher Education, 2001.

[17] Stone P J, Hunt E B. A Computer Approach to Content Analysis: Studies Using the General Inquirer System [C]. Proceedings of Spring Joint Computer Conference, ACM, 1963: 241-256.

[18] Wilson T, Wiebe J, Hoffmann P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis [C]. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, ACM, 2005: 347-354.


## 作者简介


顾文涛,浙江工商大学统计与数学学院副教授、硕士生导师。研究方向为金融统计、公司金融。

王儒,浙江工商大学统计与数学学院应用统计系硕士研究生。研究方向为文本分析。

郑肃豪(通讯作者),浙江工商大学统计与数学学院应用统计系硕士研究生。研究方向为金融计量。电子邮箱: 491088315@qq.com。

杨永伟,浙江工商大学统计与数学学院数量经济系硕士研究生。研究方向为风险管理。


(责任编辑: 董　倩)