



基于市场情绪视角的智能资产配置

弗兰克Z. 兴

计算机科学与工程学院,
新加坡南洋理工大学新加坡分校

埃里克坎布里亚

计算机科学与工程学院,
新加坡南洋理工大学新加坡分校

罗伊E. 韦尔施

麻省理工学院斯隆管理学院,
麻省理工学院, 美国剑桥

摘要 - 市场参与者的情绪指数近年来已被广泛用于股市预测。许多金融信息供应商也将其作为服务提供。但是, 很少讨论利用资产配置框架下的市场情绪。在本文中, 我们研究了市场情绪在资产配置问题中的作用。我们建议在情感分析和文本挖掘技术的帮助下, 从社交媒体计算情绪时间序列。一种新的神经网络设计, 建立在evolv-的集合之上

聚类和长期短期记忆用于将情绪信息形式化为市场观点。这些观点后来通过贝叶斯方法整合到现代投资组合理论中。我们从多个方面分析了这种资产配置模型的表现, 例如投资组合的稳定性, 情绪时间序列的计算以及模拟中的盈利能力。实验结果表明, 我们的模型优于一些最成功的预测技术。由于引入了不断发展的聚类方法, 市场观点的估计准确性得到显著提高。

数字对象标识符10.1109 / MCI.2018.2866727
出版日期: 2018年10月15日

通讯作者: Erik Cambria (电子邮件: cambria@ntu.edu.sg)

I. 介绍

金融市场是人类社会中最复杂，最混乱的动态系统之一。许多因素可能通过资产的出价和报价促成市场价格的波动。在这价格形成机制，市场参与者的心理和行为发挥着重要作用。公众情绪是一种非常有效和通用的变量，反映了市场参与者的态度。此外，Web的兴起

2. 0应用程序和社交媒体的日益普及加速了信息的传播，这更加重视了市场上的主观观点。实证研究[1]表明，主要市场的当前股价走势主要受到新信息和投资者信念的影响。

我们认为结合公众情绪对股市预测任务有利的另一个原因是这种方法带来了公共但增量的信息。相比之下，许多技术分析师完全依赖于对过去价格序列模式的挖掘。在最近将人工智能技术，特别是机器学习和深度神经网络应用于股市预测的趋势中，计算机科学界的很大一部分具有相同的局限性。然而，正如混沌理论和[2]中的许多案例所表明的那样，即使对于确定性系统，随着时间的推移，也没有“可检测的模式”。这并不一定意味着当前价格反映了有效市场假设(EMH)所暗示的所有过去信息，但由于价格受新信息驱动，过去的模式很快消失。因此，如果他们只是用过去的价格建立模型，那么模式追逐者总是落后一步。

除了价格系列，还有文献中的模型其中包括宏观经济变量，如公司的账面价值和多因素模型所建议的投资[3]。然而，这些模型的问题是这些因素的更新通常很慢。与许多经济因素不同，公众情绪可以被即时监控，并被估计为个人市场情绪的集合。以前的研究调查了各种公共情绪来源，如股票留言板[4]，微博平台[5]，报纸[6]，真正简单聚合(RSS)[7]等[8]。Wuthrich等人。[6]使用来自专家系统的事件加权关键词元组来衡量公众情绪；Zhang和Skiena [7]利用词级正面和负面计数来推导特定公司的极性和主观性；Antweiler和Frank [4]手动标记

一些消息来训练朴素贝叶斯分类器，该分类器预测看涨，看跌或两者都不基于消息的词袋表示；Smailovic'等人。[5]培训了支持向量

机器(SVM)基于由表情符号分类的大型推文数据集。最近，Weichselbraun等人。[8]提出了基于挖掘知识库的社交媒体流情感分析，丰富了依赖树。虽然使用不同

从知识工程到机器学习的技术，其中许多已经报道了公共情绪之间的相关性

和价格变动。统计检验和模拟结果也表明了公众情绪的预测能力[9, 10]。

尽管在股票市场预测中具有重要作用，但个人根据一组公共情绪数据和预测价格作出投资决策是不够或不直接的。因为公众情绪并不直接影响市场：它通过市场参与者的观点及其后续行为间接地做到了。这种互动通常被称为博弈论中的高阶信念。然后，一个自然产生的问题是关于将市民观点与公众情绪联系起来[11]。然而，在特定情况下，关于市场观点如何由公众情绪形成的机制的讨论被严重忽视。在本文中，我们将解决将公共情绪纳入资产配置框架的问题。市场观点是从情绪时间序列计算形成的，作为投资者的先验信念。交易模拟和实验证明了我们制定市场情绪观点的高质量方法。与各种最先进的资产配置策略相比，使用此先前情绪的信息增强导致平均年度投资组合收益率超过10%。

本文的其余部分安排如下：下一节本节提供了现代投资组合理论的背景，并解释了贝叶斯资产配置的概念；下面，我们描述建模市场情绪观点和优化目标；接下来，我们提出了生成情感时间序列的方法；之后，我们通过运行具有各种实验设置的交易模拟来评估我们的方法；最后，我们讨论了我们的研究结果，并提出了结论性意见。

II. 资产配置问题

A. 均值 - 方差法

半个多世纪以来，投资组合建设模式一直普遍适用于投资。鉴于可用资本总额作为约束，投资者需要将其分配给不同的资产。通常，产生更高回报的资产也承担更多风险。基于资产收益率与投资者风险之间的权衡，Markowitz提出了“有效投资组合”的概念[12, 13]。考虑一个周期模型。假设n资产被选中并且第i个被赋予权重 w_i ，那么投资组合收益将是每个资产的预期收益的加权平均值，投资组合风险可以通过收益向量的方差来衡量。因此，“有效的投资组合”符合以下条件：

$$\begin{aligned} & \text{最大化} \quad \sum_{i=1}^n w_i \mu_i - \frac{d}{2} \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij} \\ & \text{受制于} \quad w_i \geq 0 \text{ 和 } \sum_{i=1}^n w_i = 1, \quad i = 1, 2, \dots, n \quad (1) \end{aligned}$$

其中d是风险厌恶系数， μ_i 表示预期值资产i的回报， σ_{ij} 是资产i和资产数量之间的协方差。

j, n

然后，通过等式1的一阶条件给出有效投资组合的优化权重：

$$w = (dR)^{-1} n \quad (2)$$

其中 n ($n \times 1$) 是由预期回报 n_i 和组成的向量

R ($n \times n$) 是资产收益的协方差矩阵。在风险水平持有 w ，有效投资组合在所有其他替代方案中实现最大预期回报，对于预期收益等于持有 w 的所有投资组合，有效投资组合具有最小风险度量[14]。

然而， n 和 R 在实践中都是未知的。解决这个问题的传统方法是根据观察到的过去资产价格，使用他们的估计 n^t 和 R^t 。用一个长度为 T 的时间窗口，我们可以计算出两个最大似然估计：

$$\hat{n} = \frac{1}{T} \sum_{t=1}^T \frac{p_t - p_{t-1}}{p_{t-1}} \quad (3)$$

和

$$\hat{R} = \frac{1}{T} \sum_{t=1}^T \frac{p_t - p_{t-1}}{p_{t-1}} \frac{p_t - p_{t-1}}{p_{t-1}} - \hat{n} \hat{n}^T \quad (4)$$

其中 p_t 是时间点 t 的长度为 n 的价格向量。注意到价格系列在实际中通常是非平稳的

因此，估计人员对 T 的选择非常敏感。马科维茨的情况更加严重。模型本身对于作为输入的收益和波动率的估计量是不稳定的，因为误差在矩阵的乘法期间传播。因此，Markowitz模型经常提供许多零位和不平衡的投资组合[15]。

B. 贝叶斯资产配置

已经开发了许多理论方法来结合贝叶斯先验[16]。与原始Markowitz模型不同，贝叶斯视角将 n 和 R 视为固定数，而不是随机变量。人们只能推断出他们的概率分布函数(pdf)。直观地，观察到的样本大小 T 可以被包括为所谓的“扩散先验”以指示参数估计的不确定性。通过这样做，pdf对于较小的 T 将更平坦，表明更宽的置信区间和更少的样本。因此，资产在贝叶斯框架中风险更高，因为参数不确定性可能是始终存在的额外风险来源。然而，利用上述漫反射先验，优化的权重向量只是Markowitz模型的标量调整，这在信息杠杆方面几乎没有差别。为了展示贝叶斯方法的决定性优势，引出信息变量至关重要[15]，在我们的案例中，这是一种情绪先验。为了优雅地将先前的情绪与其他市场基本面结合起来，我们采用Black和Litterman提出的特定形式的贝叶斯方法[17]。在Black-Litterman模型中，投资组合收益的概率分布由两个前因推断：均衡风险溢价(Π)，以及一组关于投资者预期收益的观点。通常情况下，

均衡风险溢价按资本资产定价模型(CAPM)计算。CAPM指出，对于资产 i ，均衡风险溢价与市场溢价成正比：

$$P_i = \beta_i (\Pi + \sigma_{\Pi}^2 / \sigma_i^2) \quad (5)$$

哪里 σ_{Π} 是市场预期的回报，而且是 σ_i 没有风险利率。

Black-Litterman模型假设均衡收益正常分布为 $\text{req} \sim N(\Pi, \Sigma)$ ，其中 R 是资产收益的协方差矩阵， Σ 是 Π 的CAPM估计的置信水平的指标。市场对预期收益的看法通常也分布为 $\text{views} \sim N(Q, \Lambda)$ 。我们用 r_{BL} 表示投资组合回报的后验分布，其中下标代表Black-Litterman模型。然后，它是 -

很明显， r_{BL} 也可以写成正态分布 $N(n_{BL}, R_{BL})$ ，其中 n_{BL} 和 R_{BL} 都可以从贝叶斯定理导出：

$$pdf(n_{BL}) = \frac{pdf(\Pi) \prod_i pdf(P_i)}{pdf(\Pi)} \quad (6)$$

通过 n_{BL} 和 R_{BL} ，优化的贝叶斯投资组合具有类似于等式2的形式：

$$w_{BL} = (dR_{BL})^{-1} n_{BL} \quad (7)$$

下一节将讨论Black-Litterman模型如何以更自然和人类可理解的方式呈现市场观点，而不是明确给出贝叶斯后验收益的pdf。

III. 市场观点

从 Q 和 Λ 的物理意义出发，Black-Litterman模型定义了两种类型的市场观点[18]。相对观点采取的形式为“我有一个信心，即资产 x 将超过资产 c % a %（就预期回报而言）”；绝对观点采取的形式为“我对资产 z 的市场表现优于 b %”。因此，我们获得市场观点的定义如下。

定义1

对于由 n 个资产组成的投资组合，一组 k 个视图可以由三个矩阵 P_k ， Q_k ， 1 和 X_k ， k 表示。 P 表示视图中提到的资产。 P 的每一行的总和应该是 0 （对于相对视图）或 1 （对于绝对视图）； Q 是每个视图的预期回报；置信矩阵 Λ 是视图之间协方差的度量。

Black-Litterman模型假设视图彼此独立，因此置信矩阵可写为 $\Lambda = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$ 。按照[19]中描述的步骤，可以从等式6和定义1进一步推导出：

$$n_{BL} = [(xR)^{-1} + P_1 X^T P_1]^{-1} [(xR)^{-1} P + P_1 X^T Q] \quad (8)$$

$$RBL = R + [(xR)^{-1} + P]^{-1} \quad (9)$$

实际上，更频繁地使用更容易计算的市场视图定义，如下所示。

定义2

n 个资产的市场观点可以用三个矩阵表示

$P, n, Q, n, 1$ 和 X, n, n ，其中 P, n 是单位矩阵； Q, n, n ；

X, n, n 是非负对角矩阵。

可以在数学上证明这两个定义是

相当于表现力。然而，定义2更直观，因为可以消除矩阵 P 。并且只有当定义2成立时，我们才能使用Black-Litterman假设可以使用多元正态分布来描述视图。最后，我们的任务可以重新定义为在情绪先验的帮助下估计等式8和9中的变量。

A. 估计波动性，信心和回报

我们采用CAPM计算均衡风险溢价 P 。由此得出，Black-Litterman模型中预期投资组合收益的后验分布参数的估计取决于三个因素：均衡

librium波动率作为协方差矩阵 R ，投资者对自己观点的信心 X ，和投资者的预期在他的观点中返回 Q 。

我们的方法使用过去 k 天观察到的回报来计算协方差矩阵的后期。对于资产 i 和资产 j ，元素协方差矩阵 R 中的 v_{ij} 估计如下：

$$t = \frac{k}{N} \left(\frac{1}{N} \sum_{n=1}^k (r_i - \bar{r}_i)(r_j - \bar{r}_j) \right) / \left(\frac{1}{N} \sum_{n=1}^k (r_i - \bar{r}_i)^2 \right)^{1/2} \left(\frac{1}{N} \sum_{n=1}^k (r_j - \bar{r}_j)^2 \right)^{1/2} \quad (10)$$

其 $\bar{r}_i = \frac{1}{N} \sum_{n=1}^N r_i$

在Black-Litterman模型的最原始形式中，置信矩阵 X 根据投资者的经验手动设置。而在[18]给出的数值例子中，置信矩阵是从协方差矩阵导出的：

$$X = P \quad (11)$$

这是因为 $P(xR)P1$ 可以理解为协方差视图中预期收益的矩阵也是如此。使用定义2，更容易理解该估计，因为 P 是单位矩阵， $P(xR)P1$ 已经是对角线。基本假设是资产 i 的绝对视图的方差与资产 i 的波动率成正比。如果过去的一系列资产 i 意味着高风险，那么无论市场观点如何形成，投资者对其的信心都会降低。在这个初始情况下， X 的估计利用过去的资产价格波动信息。

预期回报与市场情绪的关系最为显着。我们的假设是存在响应

冲击市场情绪的策略，从统计上获利（产生阿尔法）。假设Black-Litterman代理使用了

过去的价格系列 (p_t, k) 和交易量 (v_t, k) 以经验形式和更新其意见的预期回报，我们进一步

使用当前时间市场对资产 (s_t) 的情绪作为先验。我们使用基于演化聚类方法 (ECM) 和长期短期记忆 (LSTM) 网络的新型深度递归神经网络 (RNN) 设计来学习这种时变策略，因此称为ECM-LSTM：

$$p_{t+1} = \frac{p_t + 1}{p_t} \quad (12)$$

ECM [20]通常用于在线系统，在该系统中，它执行一次通过，最大距离的聚类过程而无需任何优化。由于其有效记录和更新质心和聚类半径的性质，该方法非常快。LSTM是具有门控单元的特殊类型的RNN。LSTM单元通常包括输入门，遗忘门和输出门。所有门都使用当前输入和先前输出状态进行更新。据称这种单元体系结构非常适合学习预测具有未知大小滞后和长期事件依赖性的时间序列 [21, 22]。

ECM-LSTM受到观察的启发，即通过简单地应用LSTM进行的预测适应输入数据

太快。而通常是现实世界的金融时间序列非常嘈杂，这将导致过度拟合无意义的信号

如果用于离线训练。ECM机制首先被提议用于分区输入空间以学习规则

模糊推理系统。同样，我们可以赋予LSTM通过仅从关键的新传入数据中学习，即在更新旧的聚类模式时具有稳定性的模型。

[22]表明，它的变体都不能在各种任务上显着改进标准LSTM架构。因此，我们实现了[22]中描述的vanilla LSTM单元。ECM-LSTM训练和预测程序在算法1中描述，其中 v 表示S形函数，

是前状态预测而最后一个观察的事实或投资者的预期

预期回报。输入门，遗忘门和输出门的激活功能由 i ， f 和 o 表示。 W 是状态转移矩阵， b 是偏置矢量。每个LSTM单元的状态时间点 t 由其先前状态 c_{t-1} 的当前周期信息更新。 (C_i, R_i) 是聚类质心和输入向量空间的相应半径。

B. 最优市场情绪观

我们使用Black-Litterman模型的逆优化问题得出定义2中的最优市场视图 $[P, Q, X]$ 和情绪条件预期收益。考虑投资组合的后期模型，我们的目标是在期间 $(t + 1)$ 最大化资本金额：

$$\frac{capital_{alt+1}}{capital_{alt}} = \frac{p_{t+1}}{p_t} \quad (13)$$

由于 w_t 独立于 $Capital_t$ ，因此对于每个时段 t ，最佳投资组合权重为：

$$\operatorname{argmax}_W WT' = W + 1 \ 9 \ 8 \ PT \ PT \quad (14)$$

其中8和9是逐元素运算符。显然，等式14的解是一热矢量表示

其中资产的权重最大 $p_t \leftarrow -5$
等于1。解释可以没有卖空和交易费用，每天应将其全部资金再投资于下一时期增长最快的资产。让这个 w 成为
 w 在等式7中，我们将：

$$W_t = (dRBL_t)^{-1} nBL_t \quad (15)$$

用等式8和9代替RBL， t 和 nBL ， t 代表周期
我们将：

$$w_t = [d \ (rt + [(xrt)^{-1} + P]X^t)^{-1}p]^{-1} \cdot [\ (xRt)^{-1} + P]X^t)^{-1} [\ (xRt)^{-1}Pt_t + P]X^t)^{-1}Q] \quad (16)$$

因此，每个时期 t 的市场观点的最佳预期收益可以从等式16
求解：

$$\begin{aligned} Q_t(x) &= T \ (D \ tf228) \ [\ (XRT)^{-1} + (P \ tf230 \ PLX) \] \\ &\quad \# \ [rt + [(xrt)^{-1} + P]X^t)^{-1}p]^{-1}w - \\ &\quad (xrt)^{-1}Pt] \\ &= d \ [X^t \ (xRt)^{-1} + I] [Rt + \quad (17) \\ &\quad [\ (xRt)^{-1} + X^t)^{-1} \]^{-1}w \\ &\quad - X^t \ (xRt)^{-1}Pt \end{aligned}$$

IV. Sentic 计算

市场观点需要从大量文本数据中总结出情绪。情绪时间序列的质量显然是至关重要的，因为数据后来被用于估计预期收益 Q 的模型训练中。这是一项涉及其他自然语言处理技术的非平凡情感分析任务，例如命名实体识别，词极性消歧，讽刺检测 and 方面提取[23]。Sentic computing [24]是最先进的框架，不仅可以在文档或段落级别，还可以在句子，子句和概念级别上对文本进行情感分析。与统计方法相比，sentic计算结合了基于知识的极性推理和备用机器学习技术。一个基本的统计方法计算句子中的正面和负面词；但是，没有考虑句子结构。通过平均单词极性，正面和负面的单词将相互抵消，这给分析复杂情境中的情绪带来了困难。

Sentic计算主要利用概念级知识
边缘基础称为SenticNet¹ [25]，是100,000个概念的常识知识库，以及一种语义规则[26]，这是一组明确地捕捉文本中长期依赖性的语言规则。首先，使用Stanford类型依赖解析器从句子中提取多个关系元组[27]。然后，一个

语义分析器进一步提取概念。我们从SenticNet中查找概念，并触发sentic模式来处理
这些概念的关系和内在极性。如果概念不在SenticNet中，则该方法将转向分类器
通过机器学习建立。图1描绘了这个句子级别
极性检测过程。

Sentic计算具有大多数金融应用程序所需的高可解释性，并且在许多棘手的情况下都很强大。本节的其余部分提供了来自社交媒体的真实案例，其中sentic计算优于许多
其他技术。

A. 应用Sentic模式的例子

□ 示例1：我有一种感觉\$ AAPL会下降，但这是愚蠢的

这句话的预处理需要知道“\$ AAPL”指的是“Apple公司”，并在句子结尾处完成缺失期。然而，有趣的是，通过否认他自己以前的观点，演讲者实际上主张他对苹果公司的看涨情绪，并将这句话标记为正面。词袋模型将

算法1 ECM-LSTM训练和预测程序。

```

数据：传入数据流 $p, v, s$ 
结果：预期回报估计 $Q_t^t$ 
1  初始化LSTM参数 $W, b$ ；
2  如果 $C = Y_0$ 那么
3       $C_0 = (p_{t,k}, v_{t,k}, s_t)$ ；
4       $R_0 = 0$ ；
5      转到第15行；
6  其他
7       $dmin = \text{分钟}((p_{t,k}, v_{t,k},$ 
8      如果 $b \wedge Ri$ 
9      将 $(p_{t,k}, v_{t,k}, s_t)$  加到 $C_i$ 中，其中
10     转到第24行；
11 其他
12     (斯明,  $i$ ) = 分钟 $((p_{t,k},$ 
13      $v_{t,k}, s_t) - \text{词} + Ri)$ ；
14     如果 $Smin \geq 2Ri$ 那么
15         将 $(p_{t,k}, v_{t,k}, s_t)$  加入到 $C$ 中；
16          $i_t = v \ (w_i \ \$ \ [q^t \ t-1, p_{t,k}, v_{t,k},$ 
17          $s_t, + b_i)$ ；
18          $f_t = v \ (w_f \ \$ \ [q^t \ t-1, p_{t,k}, v_{t,k}, s_t, + b_f)$ ；
19         用, 更新 $W, b \ \frac{2(Q_{t-1} - Q_t)}{2\{l, f, o\}_{t-1}}$ ；
20     其他
21         将 $(p_{t,k}, v_{t,k}, s_t)$  加到 $C_i$ 中，其中
22          $Smin$ 成立；更新 $(C_i, Ri)$ ；
23     结束
24     结束
25     结束
26     结束
27 结束返回 $Q$ 
28  $t_t$ ；
    
```

¹<https://github.com/yurimalheiros/senticnetapi/>

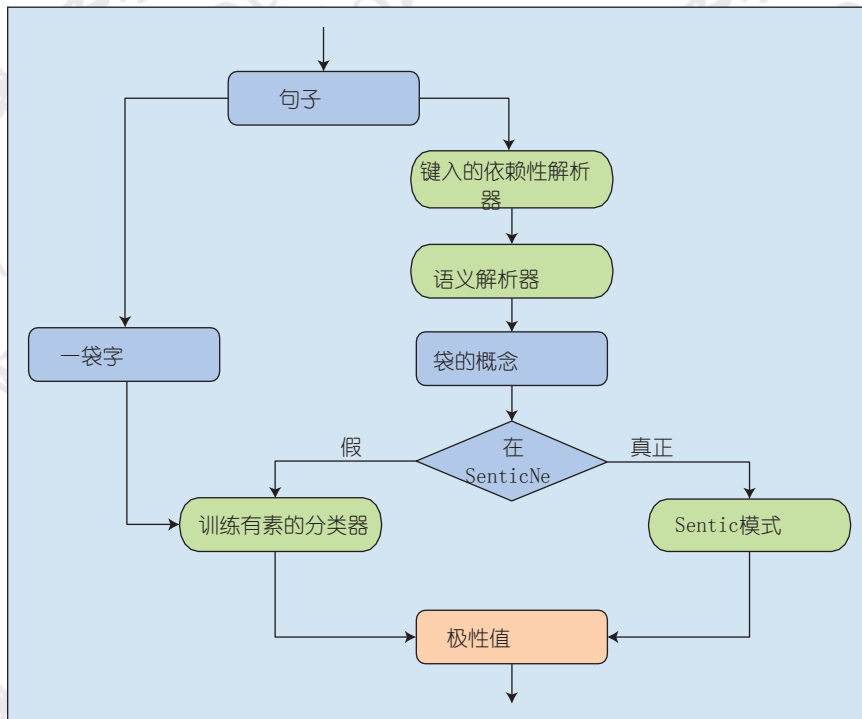


Figure 1 在句子级工作的sentic计算算法，改编自[26]。
识别“向下”和“愚蠢”，两者都是消极的并且得出结论

整个句子都是否定的。例如由Google Cloud Natural Language API (Google SA)²提供的基于机器学习的情感分析模型也未能用于该示例。Sentic计算首先会识别SenticNet中的“go_down”概念，这是负面的。这种极性将通过名义主题关系传递给“Apple公司”，而相对子句修饰符关系传递给“感觉”。请注意，这整个结构和“愚蠢”是通过相反的但是连接相关联的，因此触发了“负面但负面和正面”的正确模式，使整个句子具有正极性。

图2进一步阐述了该过程。

□ 示例2: \$ AAPL今天将再次下跌，但下跌草案正在放缓。到下周末，我认为它会被买回来

除了示例1之外，预处理还必须转换基于ASCII的撇号编码，并进行识别

两句话：“苹果今天将再次下跌，但下跌选线正在放缓。”和“到下周末我认为它会被买回来。”谷歌SA指定第一句话的情绪分数为-0.20，第二句为0.0，因此整体情绪平均为-0.10。

然而，用户将消息标记为正面，并且许多人会同意两个句子都是正面的显而易见性。Sentic计算不能为第一个句子提供正确的分数，但会将正确的整体标签视为正面。首先，but-conjunction具有最高优先级，使得第一句的极性与“down draft is slow”一致。概念“down_draft”不在SenticNet中，因此它继承了“down”的极性：-0.31。虽然股票市场的下跌趋势放缓是积极的，但在一般领域知识库SenticNet中，“is_slowing”概念是中性的。因此，负分数通过整个句子，giv-

第一句情绪评分为-0.31。这个概念

“buy_back”的情绪评分为0.82。“下一个”-0.56的情绪分数通过形容词修饰符关系传递给“下周”，并且由于“结束”和“周”之间的名词修饰符关系，“下周末”的极性被反转因此，第二句的总极性为 $1 - (1 - 0.82) (1 - 0.02) = 0.82$ 。整个消息的情绪分数为 $((0.82 + (-0.31)) / 2) = 0.26$ 。

□ 示例3: 180个呼叫赌徒的\$ AAPL静默时刻。大声笑。

此消息包含“lol”，可能是“大笑”的首字母缩写。用户通过表明有证据表明苹果的股票价格不会达到180并且贬低那些将乐观的观点视为“赌徒”的人来表达他的消极情绪。谷歌SA给出的句子的积极情绪评分为0.30，与大多数机器学习相同

基于方法，很难分析错误的位置来自。

²<https://cloud.google.com/natural-language/> [2017-12-16 访问]

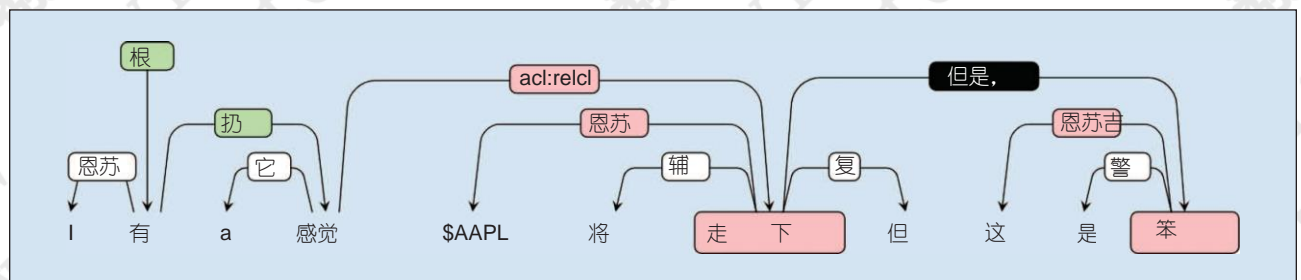


Figure 2 Sentiment得分通过依赖树从低级概念传播到句子级极性。

Sentic计算将“沉默”0.11的极性作为“沉默时刻”的情感分数，因为它是名词修饰关系。但是，案例标记“for”更新总体极性取决于后者“180名赌客”。由于“沉默的时刻”是积极的，但“赌徒”有SenticNet的负分为-0.74，这种模式引发了对“赌徒沉默的时刻”更强烈的否定性。
$$-\sqrt{|-0.74|} = -0.86.$$

通过将sentic计算应用于来自社交媒体的消息数据流，我们可以计算每日的正面和负面消息，并计算特定资产的平均情绪分数，从而形成情绪时间序列。在实验部分，我们可以观察到将消息的个体层面情绪凝聚在一起作为市场预测先验的神奇力量。

V. 实验

在本节中，我们将评估市场情绪观正式化的质量和有效性。首先，我们将sentic计算的结果与用户自己给出的标签进行比较。接下来，我们使用智能贝叶斯资产分配模型和几种投资组合构建策略的基准进行交易模拟。最后，我们讨论我们的发现。

A. 数据和情绪时间序列

在本研究中，我们收集StockTwits的意见消息，StockTwits是投资者和交易者共享财务信息的流行社交网络。此外，我们从Quandl API获取股票的历史收盘价和每日交易量；Yahoo!的市值数据金融。我们调查2017年8月14日至2017年11月16日的3个月的时间段。对于缺失值，例如周末和公众假期的收盘价，我们用最近的历史数据填补空白。

我们的数据集包括Apple的38,414条消息，Goldman Sachs的4,298条消息，辉瑞的2,157条消息，Newmont Mining的1,094条消息，星巴克的2,847条消息以及其他代码的76,553条消息。表1提供了Apple Inc. 上用户标签和sentic计算结果之间的混淆矩阵。在调查期间，总共有14,524个用户标记的消息提到Apple。从表1可以很容易地计算出极性检测的准确度为59.8%。鉴于原始数据中的噪声和我们使用的一般领域知识库，这是一个相当可接受的结果。需要注意的另一个问题是，用户标签不能完全理解为基本事实，而只是作为参考。因为无论表达的情绪如何，只有一小部分（通常少于20%）的用户会标记他们的消息，所以具有用户标签的子集可能不是无偏的样本。

但是，两个时间序列的可视化是积极的。在时间轴上扩展的负消息计数可以更好地展示两个源之间的一致性（图3）。

我们从第三个商业产品（PsychSignal）获取交易者情绪指数进行比较。他们的消息数据流收集了各种来源，包括Stocktwits和其他人。

他们的情绪分析引擎也没有透露。我们计算两个时间序列的相关性：

$$\text{相关}(s1, s2) = \frac{E((s1 - \bar{s1})(s2 - \bar{s2}))}{\sqrt{s1} \sqrt{s2}}. \quad (18)$$

表2报告了三个来源的时间序列之间的显著和正相关。

B. 交易模拟

我们通过随机选择大公司的股票来构建投资组合。该投资组合包括5只股票：Apple Inc (AAPL)，Goldman Sachs Group Inc (GS)，Pfizer Inc (PFE)，Newmont Mining Corp (NEM) 和Starbucks Corp (SBUX)。

TABLE 1 用户标签和sentic计算结果之间的混淆矩阵。

		sentic computug		总
		正	负	
你好吗？	正	7234	3748	10982
	负	2097	1445	3542
	总	9331	5193	14524

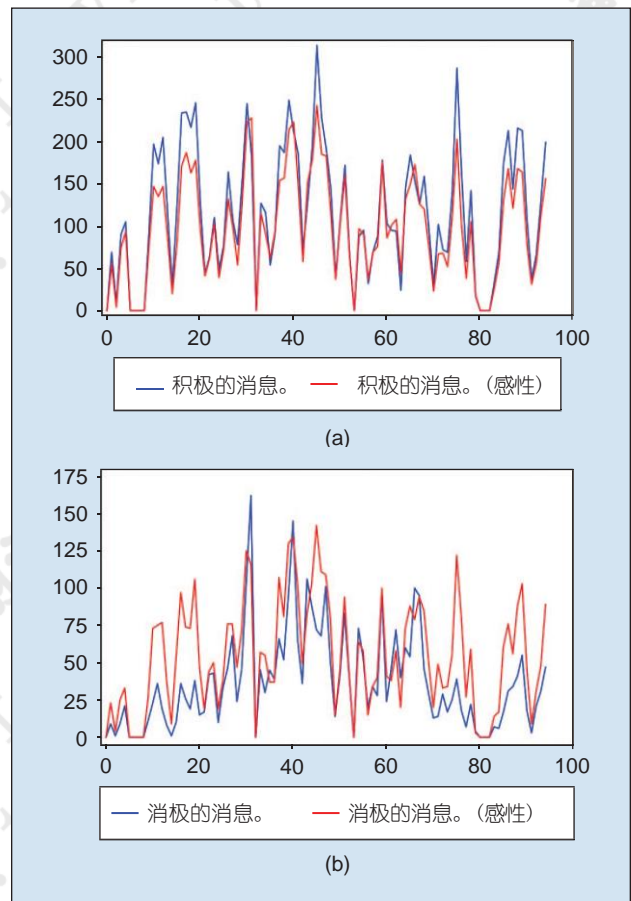


FIGURE 3 正面和负面消息的时间序列来自两个来源（x轴：天，y轴：计数）。

TABLE 2 用户标签, sentic计算和PsychSignal之间的消息时间序列的相关性。

	积极的消息	negative 信使
用户SEnTic	+0.964	+0.795
用户斗志	+0.185	+0.449
SEnTic, 斗	+0.276	+0.282

此选择涵盖纽约证券交易所和纳斯达克市场以及多元化行业, 如技术, 金融服务, 医疗保健, 非必需消费品。社交媒体发布频率在很大程度上也在这些公司中有所不同。传统行业普遍受到较少关注。每股价格根据计算所有相关变量的股票分割历史进行调整, 但不考虑股息。在模拟中, 我们假设没有卖空,

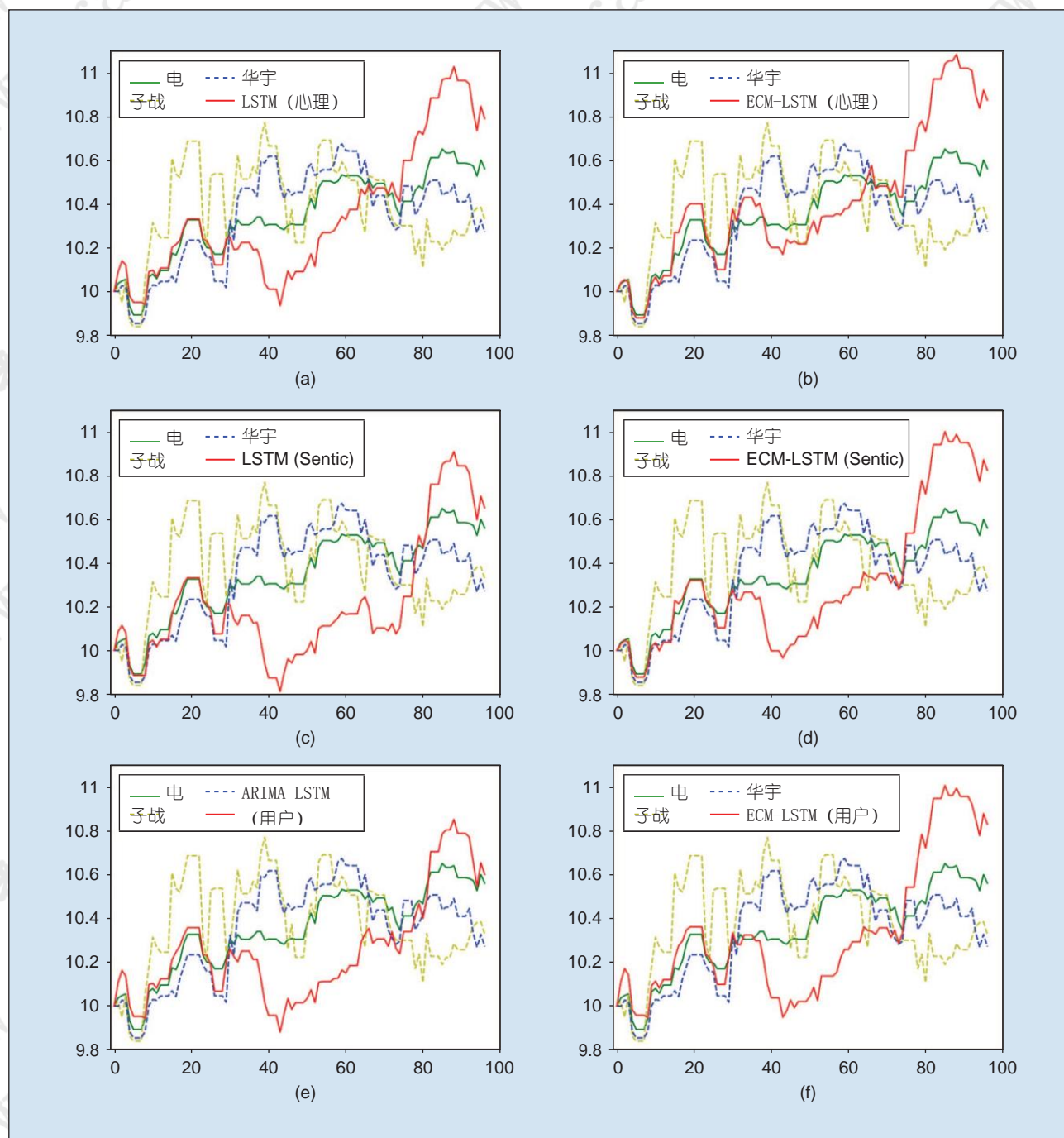


FIGURE 4 智能贝叶斯投资组合的表现, 具有不同的情绪时间序列来源 (x轴: 天, y轴: 千美元)。

税收或交易费用，我们假设投资组合投资可以无限分割，从10,000美元起。

我们将投资组合业绩与以下三种资产配置策略进行对比：

1) 等权重投资组合 (EW)：我们对整个投资组合中的五只股票持有相等的权重 (20%)

调查期间。在这种情况下，投资组合表现平均五种股票的价格变动。该策略是基本的，需要最少的信息。然而，在预测股票市场等复杂系统时，这种毫不费力的策略比许多更复杂的策略表现更好。

2) ARIMA投资组合 (ARIMA)：我们根据预测的价格每日重新投资。预测是由ARIMA (p, d, q) 模型 (自回归整数) 引入

每个库存和参数的磨碎移动平均值从历史数据推断如下。首先，增加d

直到差异化的时间序列是静止的。然后，将最大p和q设置为最后一个重要的部分自相关和自相关的顺序。最后，选择产生最小Akaike信息的 (p, q)

标准 (AIC)。实际上，除了PFE的ARIMA (0, 1, 2) 模型之外，其他价格表现出随机行走行为 (ARIMA (0, 1, 0))。

3) Holt-Winters投资组合 (HW)：我们根据一步到期的价格预测每日重新投资。预测是通过具有时变参数的Holt-Winters加法平滑方法产生的。该通过最小化模拟的均方根误差 (RMSE)，在每个时间点t指定模型HW (a^t, b^t, c^t)。

在滑动窗口中的时间序列 (t - k, t)。

请注意，ARIMA和HW组合不需要任何

然而，在此之前，它们被认为是跨越不同任务的最有效的预测技术之一

来自其他来源的信息数据不可用[28]。

我们使用Black-Litterman方法进一步构建具有来自不同来源的情绪时间序列的智能贝叶斯投资组合。根据之前的研究[18]，我们设定d = 0.25和x = 0.05。为了简化神经网络结构，我们使用单层64个LSTM单元，然后是密集连接的资产数量大小的层。隐藏的LSTM层使用20%的丢失。交易表现见图4，其中投资组合资本单位为1,000美元。

C. 评估

已经提出了多种指标来评估投资组合的绩效[29-31]。评估的难度部分是由于没有一种简单的方法来计算与黄金标准投资组合的距离。投资组合权重的相同偏差可能会因资产价格波动而放大或缩小。因此，我们放弃了指标

引入ECM机制的改进得到了以下事实的证实：就所有这些指标而言，ECM-LSTM投资组合使用相同的情绪观点来系统地表现优于同行。

权重和报告四个直接财务指标：年化回报 (AR)，夏普比率 (Sh. R)，Sortino比率 (So. R) 和最大亏损 (MDD)。

年度回报衡量给定投资组合的盈利能力。我们假设在自然年份，投资组合保持与调查期间t一样的复合增长率。设T = 365.25，我们有：

$$AR = \frac{\text{资本}^t}{\text{Capital}_0^m} \quad (19)$$

夏普比率是风险调整后的回报率。我们选择等权重投资组合为基础，因此EW的Sh. R将为1.00：

$$S. R = \frac{e^{(r_{\text{portfolio}} - r_{\text{market}})}}{V(r_{\text{portfolio}})} \quad (20)$$

Sh. R使用每日收益的标准差作为风险度量。注意，为了区分好坏风险，我们也可以使用下行回报的标准差[32]。So. R以这种方式计算。

MDD衡量投资者可能的最大损失百分比：

$$MDD = \frac{\text{最大的 } Valuet - \text{瓦鲁埃特}}{0.1 \times 1x} \quad (21)$$

具有大量MDD的资产配置策略往往会引起投资者的恐慌和不耐烦，并使投资组合面临退出风险。表3列出了这些指标。

TABLE 3前三名以粗体显示的效果指标。

	AR(%)	shr	所以 r	MDD(%)
电子战	23.07	1.00	1.00	1.76
华宇	10.72	0.56	0.61	3.79
硬件	13.03	0.34	0.36	6.16
LSTM (斗志)	33.52	0.71	0.79	3.84
LSTM(SENtic)	27.21	0.61	0.68	5.05
LSTM (用户)	24.82	0.64	0.68	4.61
ECM-LSTM(PSych)	45.51	0.74	0.82	3.45
ECM-LSTM(SENtic)	35.45	0.66	0.73	2.89
ECM-LSTM (用户)	37.53	0.71	0.87	3.40

D. 发现

图4显示, 无论用于估计市场视图的情绪来源和网络实施细节如何, 智能贝叶斯投资组合都表现出类似的移动模式。这些模式可被视为模型和投资组合选择的内在模式。此外, 在使用ECM-LSTM的三个投资组合中的两个投资组合中, 2017-09-15和2017-09-25之间其他地方发生的崩溃得到了有效纠正。

EW是实验中最稳定的策略。在表3中, EW也具有最佳的Sh. R, So. R和最小MDD。ARIMA和HW比EW更易挥发。这是因为在预测了第二天的价格之后, 整个资本被投资到唯一的获胜资产, 因此风险并没有很好地多样化。这两种策略的累积回报也不能比较这一时期的EW, 导致Sh. R和So. R非常小。这两种策略不是首选。

考虑到市场情绪的所有投资组合都获得了比上面讨论的三种基本策略更高的AR。引入ECM机制的改进得到了以下事实的证实: 就所有这些指标而言, ECM-LSTM投资组合系统性地优于同行。使用相同的情绪观点来源。

在实验中, So. Rs大于Sh. Rs, 表明这一时期的市场趋势正在上升。然而, 所有策略都有So. Rs和Sh. Rs小于1。这种观察在大多数结构良好的市场中都存在, 因为寻求更高的AR不可避免地导致投资者承担更大的单位风险。

情绪时间序列来源的质量应该是重要的, 尽管我们检查的三个来源之间的差异不是很清楚。就收集的社交媒体数据量和投资组合表现而言, PsychSignal似乎提供了最准确的情绪数据流。然而, 仅使用用户标记的消息计数有时也实现了平衡且有利的结果。

VI. 结论

市场情绪在计算智能和计量经济学界引起了极大的关注。但是, 问题通常被制定为价格预测任务而不是资产分配任务。在这项工作中, 我们提出了一种复杂的方法来从社交媒体数据流计算资产级别的市场情绪, 并使用市场视图将其整合到最先进的资产分配方法中。交叉验证实验表明, 使用sentic计算获得的情绪时间序列与某些商业工具相当。考虑到其透明性和良好的可解释性, sentic计算对于需要自然语言处理的更广泛的金融应用具有巨大潜力。

另一个重要的贡献是“在应用机器学习技术时如何处理嘈杂的财务数据”这一问题。通过引入ECM作为LSTM的筛选机制, 学习的市场观点得到了平滑, 并且有效地减少了投资组合崩溃。这种新方法将我们的资产配置策略的AR平均提高了大约10%。其他衡量标准, 如

与基于香草LSTM估计的策略相比, 夏普比率 and MDD 也得到改善。

参考

- [1] Q. Li, L. Jiang, P. Li and H. Chen, “基于张量的预测股票运动的学习”, Proc. 29th Conf. Association Advancement Artificial Intelligence, Austin, TX, USA, 2015年1月26日至30日, 第1784-1790页。
- [2] FZ Xing, E. Cambria and X. Zou, “使用模糊神经网络预测不断演化的混沌时间序列”, Proc. 联合会议Neural Networks, Anchorage, AK, USA, May 14-19, 2017, pp. 3176-3183.
- [3] EF Fama and KR French, “五因素资产定价模型”, J. Financial Econ., vol. 116, no. 1, pp. 1-22, 2015年4月。
- [4] W. Antweiler and MZ Frank, “这些谈话只是噪音吗? 互联网股票留言板的信息内容”, J. Finance, vol. 59, no. 3, pp. 1259-1294, 2004年6月。
- [5] J. Smilovic, M. Grear, N. Lavaric and M. Zidarsic, “基于递归神经网络的情绪分析”, Inf. Sci., vol. 285, pp. 181-203, 2014年11月。
- [6] B. Wutrich, V. Cho, S. Leung, D. Permutille, K. Sankaran and J. Zhang, “文本网络数据的每日股市预测”, Proc. IEEE Int. Conf. Systems Man and Cybernetics, San Diego, CA, USA, 1998年10月14日, vol. 3, pp. 2720-2725.
- [7] W. Zhang and S. Skiena, “利用博客和新闻情绪的交易策略”, Proc. 第4届国际协会进步人工智能会议Weblogs and Social Media, Washington, DC, May 23-26, 2010, pp. 375-378.
- [8] A. Weichselbraun, S. Gindl, F. Fischer, S. Vakulenko and A. Scharl, “基于方面的提取和分析来自社交媒体流的情感知识”, IEEE Intell. Syst., vol. 32, no. 3, pp. 80-88, 2017年5月。
- [9] J. Bollen, H. Mao and A. Pepe, “模拟公众情绪和情绪: Twitter情绪和社会经济现象”, Proc. 5th Int. 协会进步人工智能会议Weblogs and Social Media, Barcelona, Spain, July 17-21, 2011, pp. 450-453.
- [10] J. Bollen, H. Mao and X. Zeng, “Twitter的情绪预示着股市”, J. Com. put. Sci., vol. 2, no. 1, pp. 1-8, 2011年3月。
- [11] FZ Xing, E. Cambria, L. Malandri and C. Vercellis, “发现贝叶斯市场对智能资产配置的看法”, 在Proc. 欧洲会议机器学习与实践知识发现数据, 爱尔兰都柏林, 2018年9月10日至14日。
- [12] H. Markowitz, “投资组合选择”, J. Finance, vol. 7, pp. 77-91, 1952年3月。
- [13] MC Steinbach, “Markowitz再次访问: 金融投资组合分析中的均值-方差模型”, Soc. Ind. Appl. 数学, vol. 43, no. 1, pp. 31-85, 2001年3月。
- [14] R. Ruiz-Torres and A. Suarez, “基于基约束的投资组合选择的混合方法和降维”, IEEE Comput. INTELL. Mag., vol. 5, no. 2, pp. 92-107, 2010年4月。
- [15] D. Avramov and G. Zhou, “贝叶斯投资组合分析”, Annu. Rev. Financial Econ., vol. 2, pp. 25-47, 2010年12月。
- [16] T. Bodnar, S. Mazur and Y. Okhrin, “全球最小方差投资组合的贝叶斯估计”, Eur. J. Oper. Res., vol. 256, no. 1, pp. 292-307, 2017年1月。
- [17] F. Black and R. Litterman, “资产配置: 将投资者观点与市场均衡相结合”, J. Fixed Income, vol. 1, no. 2, pp. 7-18, 1991年7月。
- [18] G. 他和 R. Litterman, Black-Litterman模型组合背后的直觉, 1999. [在线]. 可用: <http://dx.doi.org/10.2139/ssrn.334304>. 于2018年8月27日访问。
- [19] S. Satchell and A. Scowcroft, “对Black-Litterman模型的揭秘: 管理定量和传统的投资组合构建”, J. Asset Manage., vol. 1, no. 2, pp. 138-150, 2000年9月。
- [20] NK Kasabov and Q. Song, “Denfis: 动态演化神经模糊推理系统及其在时间序列预测中的应用”, IEEE Trans. Fuzzy Syst., vol. 10, no. 2, pp. 144-154, 2002年8月。
- [21] S. Hochreiter and J. Schmidhuber, “长期短记忆”, 神经计算, 第一卷, 9, no. 8, pp. 1735-1780, 1997年11月。
- [22] K. Greff, RK Srivastava, J. Koutnik, BR Steunebrink and J. Schmidhuber, “LSTM: 搜索空间奥德赛”, IEEE Trans. 神经网络. 学习. Syst., vol. 28, no. 10, pp. 2222-2232, 2017年7月。
- [23] E. Cambria, S. Poria, A. Gelbukh and M. Thelwall, “情感分析是一个大手提箱”, IEEE Intell. Syst., vol. 32, no. 6, pp. 74-80, 2017年11月。
- [24] E. Cambria and A. Hussain, Sentic Computing: 一种基于共识的概念级情感分析框架, ser. 社会情感计算. 纽约, 纽约, 美国: 施普林格, 2015年。
- [25] E. Cambria, S. Poria, D. Hazarika and K. Kwok, “SenticNet 5: 通过上下文嵌入发现情感分析的概念原语”, Proc. 第32届 Association Advancement Artificial Intelligence Conf., New Orleans, LA, USA, Feb. 2-8, 2018, pp. 1795-1802.
- [26] S. Poria, E. Cambria, A. Gelbukh, F. Bisio and A. Hussain, “通过动态语言模式进行的情感数据流分析”, IEEE Comput. INTELL. Mag., vol. 10, no. 4, pp. 26-36, 2015年10月。
- [27] M.-C. de Marneffe and CD Manning, “斯坦福大学的类型依赖关系表示”, 在Proc. 研讨会跨框架和跨域解析器评估, 2008年, 第1-8页。
- [28] S. Makridakis and M. Hibon, “m3竞赛: 结果, 结论和影响”, Int. J. Forecasting, vol. 16, no. 4, pp. 451-476, 2000年10月。
- [29] RJ Hyndman and AB Koehler, “另一种预测准确度测量方法”, J. Forecasting, vol. 22, no. 4, pp. 679-688, 2006年10月。
- [30] M W Brandt, 第5章: 投资组合选择问题. 金融计量经济学手册. 纽约, 纽约, 美国: Elsevier, 2010.
- [31] FZ Xing, E. Cambria and RE Welsch, “基于自然语言的财务预测: 一项调查”, Artif. INTELL. Rev., vol. 50, no. 1, pp. 49-73, 2018年6月。
- [32] FA Sortino and LN Price, “下行风险框架中的业绩衡量”, J. Investing, vol. 3, no. 3, pp. 59-64, 1994年7月。