Expert Systems With Applications 135 (2019) 60–70

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

市场趋势预测的技术分析和情绪预测

安德烈亚·毕加索 a、西蒙妮·梅雷洛 a、马玉坤 b、卢卡·奥内托 a、∫、埃里克·坎布里亚 b

a 热那亚大学，Via Opera Pia 11A，热那亚 I-16145，意大利 b 南洋理工大学计算机科学与工程学院，新加坡南洋大道 50 号 e

一辆卡车
文章历史:
2018 年 11 月 9 日收到
2019 年 6 月 6 日修订
2019 年 6 月 6 日接受 2019 年 6 月 7 日在线提供
关键词:
技术分析
情感嵌入
市场趋势预测
监督学习
时间序列分析
情感分析

a b s t r a c t

股票市场预测是最具挑战性的问题之一，半个多世纪以来一直困扰着研究人员和金融分析师。为了解决这个问题，出现了两种完全相反的方法，即技术分析和基本分析。技术分析基于建立在股票价格上的数学指标进行预测，而基本面分析则利用从新闻、盈利能力和宏观经济因素中获取的信息。这些学派之间的竞争导致了许多有趣的成就，然而，迄今为止，还没有找到一个可行的解决方案。我们的工作旨在通过应用数据科学和机器学习技术将技术分析和基础分析结合起来。本文将股市预测问题映射到时间序列数据的分类任务中。技术分析的指标和新闻文章的情感都被用作输入。结果是一个稳健的预测模型，能够预测由 NASDAQ100 指数中 20 家资本最雄厚的公司组成的投资组合的趋势。为了证明我们方法的真正有效性，我们利用这些预测运行了一个高频交易模拟，达到了 80%以上的年化回报率。该项目代表着技术和基础分析相结合的一个进步，并为开发新的交易策略提供了一个起点。

1. 介绍

股票市场预测是一个具有挑战性的问题，其复杂性与可能影响价格变化的多种因素密切相关。来自不同领域的研究人员和从业者接受了挑战，因此由数学家、数据科学家、哲学家和金融分析师组成的研究单位非常普遍。这种环境的异质性导致市场理论向前迈出了重要的一步。事实上，解释市场行为的理论假说有两种:有效市场假说(EMH)和适应性市场假说(AMH)。

EMH (Fama，1991)指出，目前的市场价格充分反映了所有最近发表的新闻。这导致过去和现在的信息被立即纳入股票价格。因此，价格变化仅仅是由于新的信息或新闻，而独立于现有的信息。因为新闻不是-

本质上是可预测的，理论上，股价应该遵循随机游走模式，下一个价格的最佳赌注是当前价格。在实践中，EMH 表示，不可能“跑赢市场”，因为股票总是以其公允价值进行交易，因此，买入被低估的股票或以夸大的价格卖出股票应该是不可能的。然而，AMH (Lo，2004)试图将理性的 EMH 与非理性的行为金融原理联系起来。AMH 将进化和行为的原则应用于金融互动。行为金融学试图通过基于心理学的理论来揭示股票市场的异常。根据 AMH 的观点，利用市场效率的弱点从股票投资组合中获得正回报是可能的。

理解市场行为的另一个重要步骤是道氏理论(Rhea，1993)。它指出市场价格的变动是按趋势组织的，具体来说是三种不同的趋势，取决于

*对应作者。

电子邮件地址:andrea.picasso@smartlab.ws (A。毕加索)，simone.merello @smartlab.ws (S。Merello)，mayu0010@e.ntu.edu.sg (Y。Ma)，luca.oneto@unige.it (L。Oneto)，cambria@ntu.edu.sg (E。形成层)。

https://doi.org/10.1016/j.eswa.2019.06.014 0957-4174/2019 爱思唯尔有限公司。保留所有权利。

它们的相关性。因此，从业者开发了预测市场趋势的技术，这导致了两种不同学派的诞生:技术分析和基础分析。

技术分析师认为，价格能够详尽解释市场走势；因此，他们的策略是基于股票价格和计算出来的数学指标，比如 RSI、MACD 和布林杰波段。他们通过从烛台图中提取技术模式和探索线性方法来进行时间序列分析，如 Box-Jenkins 自回归综合移动平均线(ARIMA) (Box & Jenkins，1970)，这是时间序列预测中流行的模型之一。作为后续步骤，赫斯特指数(Hurst，1951)，一种用于对时间序列进行分类的统计度量，被证明在理解市场行为方面是有用的。事实上，正如钱和拉希德(2004)发现的，当赫斯特指数是根据价格值计算时，它提供了一种趋势可预测性的度量。

如今，随着机器学习和深度学习预测工具的力量不断增强，这一过程已经从金融分析师手中转移到数据科学家手中。张等。提出了一种基于和神经网络的时间序列预测混合模型(张，2003)和机器学习技术已在许多研究工作中应用于市场预测(胡，刘，边，刘，&刘，2018；Oancea & Ciucu，2014 年；姚，谭，& Poh，1999)。黄、中森和王(2005)利用一个由 676 对观测数据组成的小数据集，利用支持向量机()对股票市场的走势进行预测，准确率接近 70%。我们认为，增加数据集的维度可以带来更值得信赖的性能，因为小数据集限制了模型的通用化。其他研究人员在神经网络结构中输入了更大的数据集，但目标是只预测市场的特定指数(克里斯蒂亚诺尼&肖韦-泰勒，2000；姚等。1999). 姚等。(1999)在他们的工作中，开发了一个只预测单一指数的模型，吉隆坡证券交易所使用了大约 20 0 0 个样本的数据集。

在技术分析中，策略仅基于股票的价格时间序列，而在基本面分析中(Abarbanell & Bushee，1998)，交易决策是根据公司的财务状况和宏观经济指标(如息税折旧摊销前利润、市盈率、收入、股本回报率和股息收益率)做出的。因此，基本面分析师在股票内在价值高于/低于市场价格时买入/卖出股票；尽管如此，EMH 的支持者认为股票的内在价值总是等于它的当前价格(班迪，2007)。

如今，即使在这个领域，机器学习和数据科学的重要性也越来越大，这一过程的结果就是情绪分析在金融市场中的应用。情感分析旨在从文本等各种信息源中自动提取情感(李等，。2017;马，彭，&坎布里亚，2018)，图像(尤，罗，金，&杨，2015)和视频(哈扎尔卡等人，。2018;茨苓，坎布里亚，八派，侯赛因，2017)。

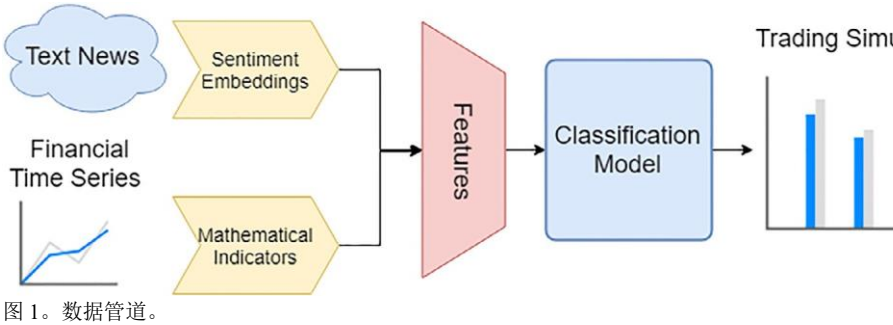


图 1。数据管道。

彭和蒋(2015)在一个神经网络模型中利用从文本中提取的情感信息作为嵌入信息来预测市场走势，事实证明这是有效的。Luss 和 D'Aspremont (2015)应用多通道学习将来自股票回报的信息与新闻文本相结合。他们观察到，虽然使用文本或返回都无法预测返回的方向，但它们的大小是预先规定的。继他们的成就之后，我们通过根据价格的相关变化的大小将测试样本分成不同的桶，沿着不同的趋势变化大小评估了我们的模型的性能。从新闻开始，各种文本来源被用来推断市场情绪和产生预测(舒马赫&陈，2006；20 09)和金融博客(德乔杜里，孙达拉姆，约翰&塞利格曼，2008；哦&盛，2011)到推文(博伦，毛，&曾，2011；米塔尔&戈尔，2012；Rao & Srivastava，2012；Si 等人，。2013).

最近，吴、苏、于和常(2012)的工作报告了在回归问题中利用新闻和技术信息时性能的提高。我们认为他们的发现是更深入调查的起点。

我们的项目旨在开发一个稳健的模型，能够通过利用来自价格时间序列和情绪的信息来预测未来的市场趋势。这将使技术分析师和基础分析师能够一起工作，提高股票市场预测的性能。

为了解决这个问题，毕加索等人。(2018 年)，我们研究了一个股票组合，这是 NASDAQ100 中列出的 20 只资本最大的股票，以避免高频交易模拟期间的流动性问题。

此外，我们相信开发更多 tickers 来证明我们方法的统计相关性的重要性。在进行时间序列预测时，经常使用不平衡的标签。遵循阿明等人的指导方针。(2016)和毕加索等人。(2018 年)，我们的方法旨在解决这个问题。从他们的工作开始，我们通过在模型选择过程中使用几何分数改进了平衡技术。我们应用了三种不同的增加复杂性的模型，即随机森林、支持向量机和前馈神经网络。一个金融基准被用来比较预测，以便在真实的交易场景中证明它们的有效性。我们的评估分为两步，因为正如邢、坎布里亚和韦尔施(2018)所述，当应用于金融领域时，数据科学模型的评估可能很困难。第一步考虑了与机器学习领域更相关的指标，以了解模型的统计行为。在第二步中，我们通过来自交易模拟的结果来评估模型，该模拟基于模型预测、使用收益、最大压降和夏普比率作为业绩指标。两步评估的使用使我们的方法能够克服与以前工作相关的问题和偏见。第一步，证明了该分类器的有效性。通过第二步，我们预测的真正有效性得到了证实。我们项目的整个流程如图 1 所示。1 使实验流程更清晰。在实验阶段，我们的模型被证明是有效的，从统计和财务评估，实现了 85.2%的投资组合的年化回报。

论文的其余部分组织如下:第二部分介绍研究问题；第 3 节强调了我们处理这一问题的新颖之处；第 4 节描述了可用的数据集；第 5 部分报告了我们每个实验的设置和结果；最后，第六部分指出了本文的结论和未来的工作。

2. 问题形式化

预测市场趋势需要解决的统计问题是一个自动递归分类问题。类别 X 的输入被定义为向量序列:X = [x (0)，x (1)，..,。**x(n1)，x (n)]**

其中 n 表示样本的数量。通过选择一个通用样本 x (t) ∈ R F，其中 F 是特征的数量，t 是样本的时间戳，它可以分解为:

x (t) = [x (t) 0，x (t) 1，.。 x (t) F]

分类问题的目标被定义为一系列标签:

y = [y (0)，y (1)，.。 y (n)]

X 的长度相同，使得 y ∈ {0，1}的每个元素。

在 y (t)的计算中，我们引入了超参数 w，它确定了待分类趋势的大小。因此，给定输入 X，分类过程旨在区分两个类别:与 y (t) = 1 相关联的正类别和与 y (t) = 0 相关联的负类别。特别是，我们认为正样本表示在时间 t 和 t + w 之间收盘价 pc 的增加，而负样本则相反。

如果负趋势在[t，t + w]，

简单移动平均线 普通

指数移动平均线 均线(t1))*多+均线(t1)多 ，= timeperiodEMA

相对强度指数 1+100 RS(t)，RS(t) = AA v v gGain gLoss ((tt))

布林带 上限值(t)= 20 ∫ 形状记忆合金(t，N)+(40 ∫ 标准(pc))中限值(t)= 20 ∫ 形状记忆合金(t，N)下限值(t)= 20 ∫ 形状记忆合金(t)(40 ∫ 标准(pc))

随机振荡器 KDJ

真实范围

平均真实射程

WR 指标

pc =收盘价，po =开盘价，pl =低价，ph =高价，std =标准差。n.

$$\begin{cases} y(t) = 0 \\ \quad \text{如果[t，t + w]中出现正趋势，则 y (t) = 1} \end{cases}$$

现在报道了计算解释的标号所涉及的数学。阶跃函数 1 应用于价格增量，将结果从 R → { 0，1 }移动，因此 y (t)计算如下:

$y(t)= 1(PC(t+w)-PC(t))$

其中 pc (t)表示所选股票在时间 t 的收盘价，w 表示待预测趋势的长度。

3. 方法学

3.1。 数据预处理

在这项研究工作中，以前解释的自动递归分类问题被用作预测未来市场趋势的方法。属于输入序列 X 的向量 X 的每个分量 X 被归一化，使得 $x \in \mathbb{R} \rightarrow x' \in [0, 1]$加速收敛。特别是，最小-最大规范是在阿尔·沙拉比、沙班和卡斯-贝(2006)的建议下应用的。最小-最大归一化是一种简单的技术，其中数据可以被推入预定义的边界[C，D] (Patro & Sahu，2015)。

$$x' = \frac{x - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})} * (D - C) + C$$

其中x' denotes the Min-Max归一化 x 值。分类任务中用作输入的特征集来自两个不同的来源，即文本和数字来源。文本来源由关于研究中的股票的新闻组成，而数字来源指的是它的价格。在整个实验过程中，对数字源(“价格”)、数字源(“新闻”)及其联合(“价格与新闻”)的性能进行了评估和比较。这些特征组合在一起，最大限度地利用了新闻中包含的情绪以及从计算价格的数学指标中检索到的交易信号。从文本数据中，使用 Loughran 和 McDon- ald (2011) (L&Mc)和 affectivspace(Cambria，Fu，Bisio 和 Po- ria，2015)的字典提取了两组不同的特征。在这两种情况下，新闻都被转换成了即时嵌入。《拉夫兰和麦当劳词典》专门针对金融应用，包含不同的约束词、诉讼词、否定词、肯定词、不确定词、多余词和有趣词。之所以选择它，是因为如 Loughran 和 Mc- Donald (2011)所述，为其他学科开发的词典对金融文本中的常用词进行了错误分类。此外，在金融预测领域的许多研究论文(金等，。2013;李，谢，陈，王，&邓，2014)。另一方面，情感空间是一个向量空间模型，通过随机投影建立，允许对自然语言概念进行类比推理。在情感空间中，通过情感常识知识的维度还原，每个概念被映射到一个 100 维向量。这个过程使得与概念相关联的语义特征被一般化，因此，允许概念根据它们的语义和情感关联性被直观地聚类。这种情感输入(因为它不依赖于显式特征，而是依赖于隐式类比)能够推断多词表达所传达的情感和概率，从而实现有效的概念级情感分析(Cambria 等人，。2015). 即使影响空间不是金融领域特有的，它被选中也是因为它能够从像新闻这样的结构化文本中提取概念层面的情感。这两种方法都被用于从新闻摘要文本中提取情感嵌入。使用拉夫兰和麦克唐纳词典时，每个新闻的嵌入由词典中的约束词、诉讼词、否定词、肯定词、不确定词、多余词和有趣词组成，这些词位于新闻摘要中。因此，在一刻钟的同一时间段发布的新闻嵌入被分组在一起，并且该时间段的代表性嵌入被计算为它们之间的平均值。还包括一个额外的功能，代表在该槽中找到的新闻数量(8 个功能)。最后，获得了新闻相关特征向量，该向量是实际样本分别与窗口大小为 5、10、15、20 个时隙的先前样本的移动平均线的连接。这一过程的目标是考虑过去时间的影响，产生一个 40(8⌈5)维的向量，传达所研究的特定股票的情绪信息。一个简单的过程是用通过 SenticNet API 获得的嵌入来完成的。具体来说，从新闻中提取概念，从情感空间中检索每个概念的表示作为 100 维向量，并且计算新闻的嵌入作为其概念表示内的平均值。最终，与拉夫兰和麦克唐纳数据特征一样，一个槽的嵌入被获得作为归属新闻的嵌入的平均值，并且槽表示被获得作为先前槽的平均值(500 维向量)的连接。根据谷歌金融应用编程接口以一刻钟的频率(与新闻时段的时间一致)检索的价格数据，使用 Stockstats 库计算了不同的技术指标，它们是

与价格值连接。结果是一个 111 维的向量，由价格值和所选的一组指标串联而成。要计算的指标与毕加索等人使用的指标一致。(2018 年)，它们是根据以前的研究工作选择的(Choudhry & Garg，2008 年；黄，杨，庄，2008；金和韩，2000；水野彩香，科萨卡，亚吉马和科莫达，1998 年)。他们的数学公式见表 1。

分类任务属于时间序列预测问题，通常情况下，它会受到标签不平衡的影响，这是由价格走势趋势引起的。为了解决数据集中的这一弱点，采用了一种适当的平衡技术来避免开发一个有偏差的分类器，该分类器只能预测与训练和验证集中所表示的趋势相似的趋势。合成少数过采样技术(SMOTE)是特别选择的，因为它已被证明是从以前的研究工作中最有效的(Amin 等人，。2016;查瓦拉，鲍耶，霍尔和凯格尔迈耶，2002 年；苏露莎&布拉格斯，2010)。SMOTE 是一种过采样方法，通过创建“合成”示例对少数族裔进行过采样(Chawla 等人，。2002). 平衡技术被分别应用于列车和验证集，以使它们的标签分别平衡。测试集是不平衡的，因为要预测的趋势代表了价格的未来值，而这些值是无法修改的。

时间序列预测任务中常见的另一个问题是数据集内样本之间的依赖性。事实上，当像简单移动平均线(SMA)这样的数学指标(有关数学细节，请参考表 1)在时间 t 用 D 元素的窗口对价格进行计算时，两个相邻输入 x (t)和 x (t + 1)不再独立。在计算形状记忆合金特征(定义为 x (t) i)的过程中，考虑的收盘价值集 pc (t)仅在一个元素上不同。为了说明这个概念，x (t) i 和 x (t + 1) i 的计算如下:

Country	2004		2007		2010		2013		2016		2019	
	Amount	%	Amount	%	Amount	%	Amount	%	Amount	%	**Amount**	**%**
Argentina	1	0.0	1	0.0	2	0.0	1	0.0	1	0.0	**2**	**0.0**
Australia	107	4.1	176	4.1	192	3.8	182	2.7	121	1.9	**119**	**1.4**
Austria	15	0.6	19	0.4	20	0.4	15	0.2	19	0.3	**16**	**0.2**
Bahrain	3	0.1	3	0.1	5	0.1	9	0.1	6	0.1	**2**	**0.0**
Belgium	21	0.8	50	1.2	33	0.6	22	0.3	23	0.4	**36**	**0.4**
Brazil	4	0.1	6	0.1	14	0.3	17	0.3	20	0.3	**19**	**0.2**
Bulgaria	...	..	1	0.0	1	0.0	2	0.0	2	0.0	**2**	**0.0**
Canada	59	2.3	64	1.5	62	1.2	65	1.0	86	1.3	**109**	**1.3**
Chile	2	0.1	4	0.1	6	0.1	12	0.2	7	0.1	**8**	**0.1**
China	1	0.0	9	0.2	20	0.4	44	0.7	73	1.1	**136**	**1.6**
Chinese Taipei	9	0.4	16	0.4	18	0.4	26	0.4	27	0.4	**30**	**0.4**
Colombia	1	0.0	2	0.0	3	0.1	3	0.0	4	0.1	**4**	**0.0**
Czech Republic	2	0.1	5	0.1	5	0.1	5	0.1	4	0.1	**7**	**0.1**
Denmark	42	1.6	88	2.1	120	2.4	117	1.8	101	1.5	**63**	**0.8**
Estonia	0	0.0	1	0.0	1	0.0	0	0.0	..	..	...	.
Finland	2	0.1	8	0.2	31	0.6	15	0.2	14	0.2	**7**	**0.1**
France	67	2.6	127	3.0	152	3.0	190	2.8	181	2.8	**167**	**2.0**
Germany	120	4.6	101	2.4	109	2.2	111	1.7	116	1.8	**124**	**1.5**
Greece	4	0.2	5	0.1	5	0.1	3	0.0	1	0.0	**1**	**0.0**

x (t) i = k 1 t k)

(1)

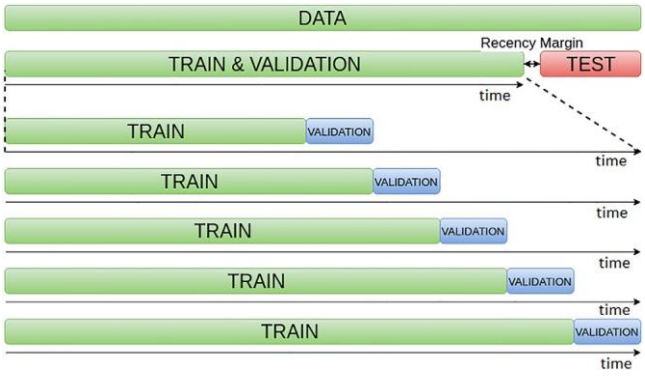
D

x (t + 1) i = k 1 k)

(2)

D

因为重叠的集合阻止了 k-fold 交叉验证的使用，因此不可能混洗数据并随机挑选训练和验证部分。否则，来自验证集的样本将强烈依赖于训练样本。为了解决这个问题，毕加索等人采用了一种特殊的交叉验证技术，称为“增加窗口交叉验证”。(2018) . 用这种技术执行的数据集分割如图 1 所示。2 .



图。2.递增窗口交叉验证。

该技术在时间序列问题的交叉验证阶段被证明是有效的。首先随着时间的推移对数据进行分类，随后从测试集中分离出训练和验证阶段，以便测试集代表训练集的“未来”。训练阶段被分成不同的折叠，在每个折叠中，训练部分被增加，并且验证集被及时向前移动。结果是一个基于折叠的训练过程，但没有打乱样本。此外，“增加窗口交叉验证”技术在训练验证和测试集之间采用了一个余量，以克服姚和 Poh (1995)强调的新近性问题。在数据集上执行分割后，可用样本被视为不同分类模型的输入，以便在它们之间进行比较。具体来说，采用了随机森林、SVM 和前馈神经网络。在下一小节中，对应用模型的简要描述据报道对实验阶段有更好的理解。

3.2。 模型

实验的第一个模型是一个射频，以便有一个基准，并获得对趋势预测任务的一些见解。射频是一种用于分类、回归和其他任务的集成学习方法，通过在训练时构建大量决策树并预测代表类模式的类来操作。决策树是一种结构，其中每个节点代表一个特征，每个链接代表一个决策，每个叶子代表一个标签。在这项工作中，射频被修剪通过基尼杂质指标。基尼不纯度是一种度量，用于衡量从集合中随机选择的元素被错误标记的频率，如果它是根据子集内的标签分布被随机标记的话。为了计算具有 J 类的一组项目的基尼杂质，假设 I∞{ 1，2，. . . ，J}，并且让 p i 是集合中用 I 类标记的项的分数。

$$(p) = 1 \quad \sum_{i=1}^{J} p^2 I G \quad - \quad 我$$

射频虽然简单，但选择它是因为它在应用于分类任务时具有显著的性能(卡伊德，萨哈和戴，2016；Kumar & Thenmozhi，2006)。

作为第二步，在趋势预测任务中使用了 SVM(cristiianini & Shawe-Taylor，20 0 0)，特别是使用了具有高斯核的 SVM(KSVM)，这在以前的研究工作中取得了有趣的性能(Choudhry & Garg，2008；黄等。2005). SVM 分类器也被称为最大边际分类器，因为它试图最大化决策界限之间的边际。通过优化以下目标函数来创建 SVM 分类器:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y(i)(\mathbf{w} * \mathbf{x}(i) - b)\right] + \lambda||\mathbf{w}||^2$$

其中 w 表示权重，b 表示偏差。目标的第一部分是一个铰链损失函数，其中 y i 是真实的标签(-1 或 1 在我们的环境中)(Gentile & Warmuth，1999)，而第二项是控制标记维数的 L-2 范数。

除了仅执行线性分类之外，SVM 还可以使用非线性核函数来有效地执行非线性分类，该非线性核函数用于将输入 x 映射到新的特征空间 x → x;。SVM 比其他分类模型表现得更好，因为它旨在最小化结构风险，而替代技术基于经验风险的最小化。换句话说，SVM 寻求最小化泛化误差的上

限，而不是最小化训练误差。因此，它不太容易受到过度拟合问题的影响。此外，优化问题的解是唯一的，不存在局部极小值。

作为最后一步，通过使用前馈神经网络实验了次符号方法。神经网络是由一堆不同的层组成的模型。层(神经元)的每个单元对其输入的线性组合应用非线性。神经元本身并不强大，但是，一旦更多的神经元在更多的层中结合在一起，网络就变成了一个极其强大的分类器(Ruck，Rogers，Kabrisky，Oxley & Suter，1990)和回归器(Cybenko，1989)。训练是通过使用反向传播算法来执行的(鲁梅尔哈特，辛顿，威廉姆斯，1986)，目标是最小化一个预定义的目标函数，这个过程不一定能达到一个最佳最小值，但它已经被经验证明达到了目标的最佳局部最小值之一。在这项工作中，建议的网络，其结构见图。3、分为四层。不同层次的神经元数量在减少；从第一个有 ne 神经，第二个有 Ne2，第三个有 Ne4，到最后一个只有一个神经元。$\prime(y)$亚当优化器(金玛&巴，2014)被用来最小化二元交叉熵 H 和 y 目标:

$$(y') = -\sum_{i=1}^{n}(y_i * log(y_i') + (1 - y_i) * log(1 -$$

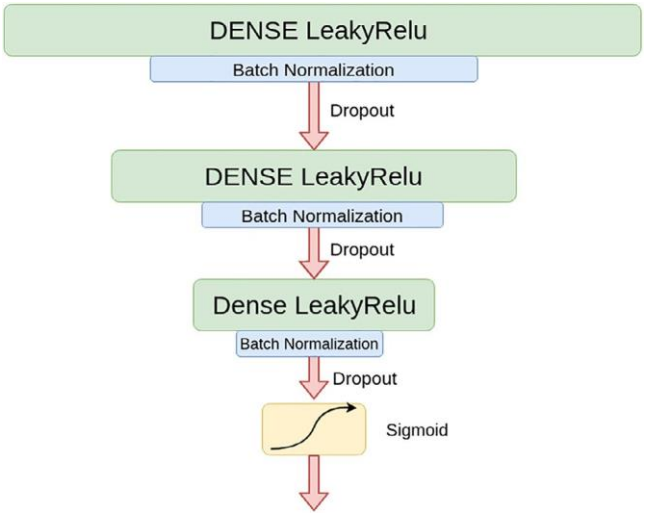
$$, \qquad \text{和}))-$$

世

其中 y i 和 y i 分别表示预测和真实标签。

在每一层中，插入一个批量归一化(Ioffe & Szegedy，2015)，以避免协变量移位(Shimodaira，20 0 0)并加快训练阶段。此外，除了最后一层使用 sigmoid 函数执行分类任务之外，每一层都采用了 Leaky Relu 作为激活函数。

为了避免过度拟合和提高网络的泛化能力，Dropout(斯里瓦斯塔瓦，辛顿，克里哲夫斯基，苏斯克弗，&



图。3.前馈神经网络结构。

Salakhutdinov，2014)应用于内部/隐藏输出。由于它具有增加权重 w 值的副作用，按照 Srivastava 等人的建议，。执行了两种不同类型正则化。特别地，最大范数正则化(Srebro，Rennie & Jaakkola，2005)被应用于权重 w，并且 L2 范数正则化被执行于密集层的隐藏输出。网络结构的微调是通过观察学习曲线进行的，最终结构见图。3 .

3.3。 模型选择和评估

这项工作的目标是开发一种能够对股票组合进行趋势预测的方法。因此，所考虑的每一个结果和指标都被计算为属于所研究的 portfo- lio 的每一个出票人的结果的平均值。在训练和验证阶段，按照前面解释的交叉验证技术，每个模型被训练更多次，并且实验不同的参数。随后，执行模型选择步骤以根据训练和验证集采用最佳参数。在说明模型选择之前，重要的是要记住，不平衡标签的存在是市场时间序列预测的一个典型问题。事实上，研究期间的市场大多是积极的(看涨)或消极的(看跌)。因此，这一阶段的主要目标是选择一个模型，该模型能够对正样本和负样本做出正确的预测，并且不仅能够实现高方向精度(市场方向的正确分类:正 y = 1 或负 y = 0)。为了达到这个目标，在模型选择期间，引入了几何分数度量。几何分数，也称为 G 分数，考虑了正反两方面的回忆，平衡了它们之间的比例。为了更好地理解它的力量，报告了 G 分数计算:

$$g\ 分数 = \sqrt{TPR[TNR}$$

如公式所示，要实现高 G 值指标，TPR、真正率和 TNR(真负率)必须高且平衡。因此，在这一阶段选择的模型能够在看涨和看跌市场做出正确的预测。总的来说，采用这个度量来提高我们的模型在正趋势和负趋势上的通用性和预测能力，消除了来自训练集分布的偏差。

表 2
在测试集上计算的用于评估的桶的详细信息。

水桶	最小值	最大值	样本数目	不平衡位置负数
一	0%	2.0%	9217 (45.4%)	55.4%
2	2.0%	4.1%	6009 (29.6%)	73.0%
3	4.1%	6.1%	3028 (14.9%)	73.6%
四	6.1%	8.1%	1384 (6.8%)	68.6%
5	8.1%	10.2%	654 (3.3%)	63.4%

股票的价值是平均的。最小值、最大值:每个时段的市场波动边界。

一旦开发了预测模型，就必须对样本外数据进行评估，以了解其真实性能。正如 Frank 等人指出的，。在评估应用于金融界的数据科学模型时，仅考虑统计指标并不全面(Xing 等人，。2018). 我们的建议是一个综合和创新的评估阶段，分为两个步骤。第一个涉及数据科学度量的计算，即方向准确度和召回值，但是它们的计算方法是原创的。事实上，在计算度量值之前，属于测试集的标签被分成五个桶。第一个桶装满了与代表市场价格最小变化的标签相关的输入。另外四个桶装满了代表市场价格百分比变化的样品。表 2 中报告了属于每个铲斗的元件的更多细节。

这个“桶化”过程的目标是了解模型在不同预测趋势下的表现。事实上，在“最相关”的样本上实现高性能，即，。代表市场价格巨大变化的数据比在不太相关的样本上获得同样的结果更有价值。这是因为，从一个交易者的角度来看，来自价格中的一个小增量的再转向经常被交易成本抵消，而市场价格中最大的增量是非常有价值的，它们产生正回报，即使在扣除交易成本之后。特别是，按照随后的规则以线性方式分割样本。定义 I = PC(t I+w)-PC(t I)样本 I 和max，Δ_{min}的市场变化分别为测试集的最高和最低价格增量；样本 I 属于 j ∈ N 的桶 j∈1；5] 如果:

$$步长 * (j-1) < \Delta_i \leq 步长 \sum (j+1)$$

其中，使用五个桶，步骤定义为:

$$步长 = \frac{max \quad \Delta_{min} -}{5}$$

因此，在模型评估的第一步中，为每个桶计算准确度和召回值。在执行第一步时，第二步是运行一个交易模拟，以检验模型生成的预测的性能。

交易策略是根据神经网络输出上的交易阈值 TSH 制定的，该阈值 TSH 基于通过观察所产生的预测而获得的精神发现。

业务单位如果预测> 0。 5 + T SH，

如果预测值< 0，则卖出。 5T SH

通过使用阈值，我们的目标是只有当我们的模型预测有一定的可信度时才进行交易，通过 sigmoid 输出值来量化，并且预测的样本代表市场价格的显著变化。事实上，当在代表市场价格微小变化的样品上交易时，收益被交易成本抵消。

现在解释确定阈值的过程。首先，对验证集的预测进行绘图，以了解它们的分布。随后，根据以下计算，阈值的值被设置为仅在预测足够“可靠”时进行交易。假设预测的分布为概率质量函数(离散概率分布)p(i) : N → N 在区间[0，1]内分成 100 个大小为 0.01 的箱，nv 为验证样本数，则 T SH ∈ R 计算如下:

$$\text{镍}_{100}^{ni} \quad s.t. \quad \sum(p(i) + p(100 -$$

$$\_\_\_\_ T\,SH = / \qquad\qquad i)) = 0。75n * v$$

$$i = 1$$

其中 ni 用作概率质量函数中的整数指数。这个计算允许我们考虑 75%的预测，根据乙状结肠输出，这些预测是最“可靠”的。

BackTrader python 库用于运行模拟，年化回报率、夏普比率和最大提款被计算为性能指标。年化回报率(aR)是一项投资在给定时间内每年赚取的几何平均金额。它以几何平均值的形式计算，以显示如果年回报率为复合收益，投资者在一段时间内将获得的收益:

$$= (1 + R)^{\frac{365}{d}} - 1$$

阿肯色州

其中 R 表示所考虑的 d 天期间的累计回报。年化回报率只是投资业绩的一个快照，并没有给投资者任何波动的迹象。因此，为了了解投资组合价值的波动性，利用了夏普比率和最大提款。夏普比率 Sr 是单位波动率或总风险的无风险率过程中获得的平均回报，由投资组合标准偏差 σ p 表示。

$$\_\_\_\_$$

$$r\ r\ Sr = \sigma p$$

从收益 R 中减去无风险利率 R，可以分离出与冒险活动相关的绩效。这种计算的一个直觉是，从事“零风险”投资的投资组合，如购买美国国债，其预期回报是无风险利率，其夏普比率正好为零。另一方面，最大压降(MDD)是指在达到新的峰值之前，从峰到峰谷的最大损失。MDD 是特定时期内下行风险的指标。

两步评估的使用使我们的方法能够克服与以前的工作相关的问题和偏见。事实上，通过第一步，所提出的分类器的能力被证明，并且通过第二步，我们的预测的真正有效性被证明。

4. 现有数据

用于进行市场趋势预测的数据集涉及 NASDAQ100 指数。它由 20 个标签组成，使这项研究的结果具有统计学意义。特别是，在交易模拟过程中，为了避免流动性问题，选择了 20 只资本化程度最高的股票。数据集的时间跨度为 2017 年 7 月 3 日至 2018 年 6 月 14 日，表 3 报告了相关资本化的 20 只在研股票的清单。对于每个股票代码，与特定股票相关的新闻都是从 Intrinio API 中检索的，而价格值的时间序列(频率为 15 分钟)是从谷歌金融 API 中下载的。由于免费版 Intrinio API 的限制，我们数据集的时间跨度是有限的。因此，即使

表 3
正在研究的股票。

股票	心脏	资本化	增益
苹果公司。	苹果公司	1026.62 亿美元	32.96%
Amazon.com 公司	亚马逊公司	923.45 亿美元	80.76%
谷歌	框架	868.65 亿美元	28.20%
微软公司	微软公司	830.73 亿美元	48.78%
脸书	运货单(freight bill)	519.84 亿美元	32.59%
英特尔公司	色调调整中心数据	225.78 亿美元	65.99%
思科系统公司。	思科系统	205.75 亿美元	42.64%
康卡斯特公司	康卡斯特	160.55 亿美元	−15.75%
百事公司	精力	155.46 亿美元	−8.73%
英伟达公司	英伟达		91.57%
网飞公司	奈飞公司	148.63 亿美元	168.78%
安进公司	安进公司	126.95 亿美元	7.52%
Adobe 系统公司	奥多比系统	1.245 亿美元	86.47%
德州仪器公司	德州仪器	107.12 亿美元	49.20%
贝宝控股公司	股票代码	102.99 亿美元	61.93%
贾勒德科学公司	协会	100.01 亿美元	0.64%
好市多批发公司	费用	963 亿美元	29.15%
高通公司	高通公司	$95.26	8.21%
博通公司	AVGO	91.12 亿美元	16.84%
预订控股公司	BKNG	89.33 亿美元	15.19%

收益:实验期间股票价值的波动。

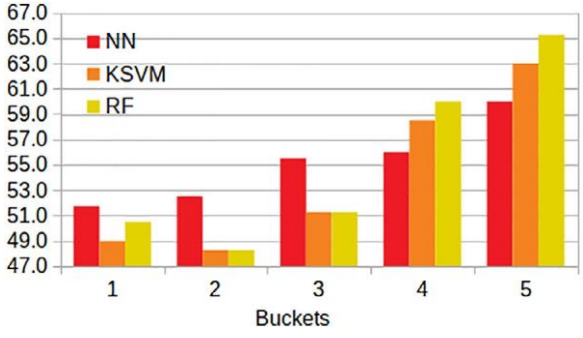
金融行业并不缺乏价格历史数据，我们的实验因为新闻数据源而被限制在报道期内。

谷歌金融的数据由开盘价、收盘价、中间价、最高价、最低价和成交量组成，没有遗漏样本。另一方面，Intrinio API 的数据集包括来自路透社或彭博等新闻机构的新闻，发布频率不像价格那么规律，因此可能会丢失样本。为了使数据集连续，在这 15 分钟内没有发布新闻的情况下，代表前 15 分钟时段新闻的样本将复制到下一个时段。这个过程遵循经验法则，即如果没有新的信息，情绪会一直持续下去。由于价格值仅在市场开盘时可用，所以在交易时段结束后发布的隔夜消息会在下一个交易日的第一个时段崩溃。嵌入向量(在第 3 节中解释)是在可用的数据集上构造的，随后作为输入输入到模型中。

5. 实验讨论

在实验阶段，开发了我们的工作方法，并使用 Python 实现了模型。具体来说，射频和 SVM 实现是从 Sklearn 检索的，而神经网络结构是使用 Keras 和 Tensorflow 作为后端构建的。创建了三个不同的数据集。第一种是由从价格中提取的特征和在此基础上计算的数学指标构成的。第二个由从新闻中提取的情感嵌入组成。最后，第三种方法是通过特征向量的连接将前两种方法结合起来。

在趋势分类任务中，要预测的趋势 w 的长度被设置为 140 个时间步长，在 15 分钟的频率数据中，这表示大约一周。这个值是根据我们以前的实验和文献的发现选择的(Merello，Picasso，Oneto & Cambria，2019；泰和曹，2001；Thomason，1999 年；徐&科恩，2018)。如表 2 所示，在实验的时间跨度内，每个时段的正/负样本百分比在 55%至 74%之间，因此市场的整体趋势是看涨的。为了减少这种不平衡因素，考虑使用适当的平衡技术。在预处理阶段



图。4.L&Mc 字典的准确性。

与每个模型一样，训练和验证集通过使用 SMOTE 算法进行平衡，邻居数量设置为 5，遵循苏露莎和布拉格斯(2010)以及 Chawla 等人已经引入的设置。(2002) . 最后，对整个输入数据集进行最小-最大归一化。

在这个初步阶段之后，拉夫兰和麦当劳字典的组合数据集(价格和新闻)被输入到三个不同的模型以及后续设置中，以初步了解我们方法的有效性。= 1在射频中，最优特征 SVM 分裂组合的计算是基于基尼系数的，而具有高斯核的是通过搜索 C 和 γ in 的值来训练的 { 10 −4 , 10 −3 . 5 . . . , 10 6 } . 相反，用神经网络探索的超参数是:

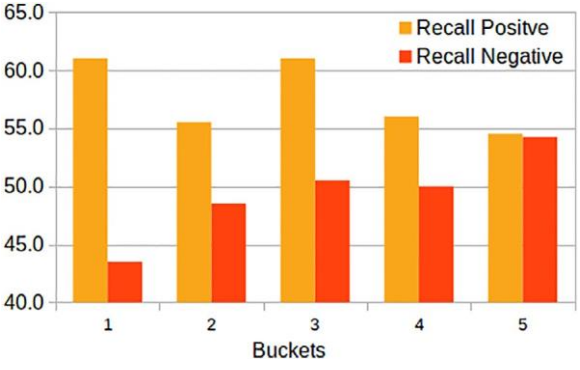
- 活动正规化 L2 0.01
- max_value = 1 的核正则化最大范数
- 辍学率 0.5。
- 第一层神经元数目 Ne = [64，128，256]
- 学习率[0.001，0.0001]
- Adam 用作优化器，其= 108，β1 = 0。9 , β2 = 0 . 金马和巴建议的 999(2014)

在训练和验证阶段之后，在样本外数据上对模型进行测试，沿着 buck- ets 的准确度值如图 2 所示。4 .

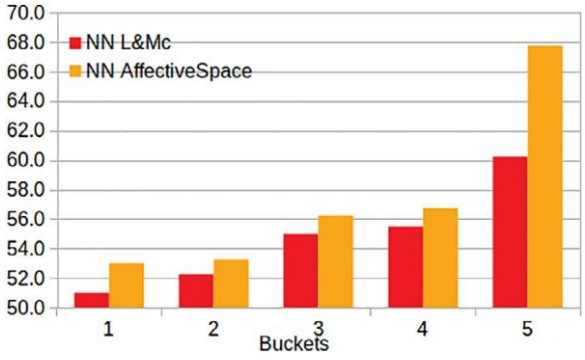
在首字母桶中，我们的神经网络的性能克服了使用射频和 KSVM 获得的结果。相反，在最后几个桶中，射频和 KSVM 的精度值更高，但是一旦射频和 KSVM 的召回值被计算出来，它们就会产生偏差。事实上，阳性和阴性样本的召回率相差超过 20%；准确地说，在射频中，阳性和阴性召回率分别为 69%和 38%。这是分类器泛化的一个问题，必须考虑到这一点，这使得准确性的值不可信。相反，神经网络召回值更加平衡，因此，为了正确评估神经网络分类器的行为，其召回值绘制在图 2 中。5 .

这证明了我们的神经网络模型能够公平地对正样本和负样本进行分类。事实上，橙色和红色图表之间的差值，分别为正和负召回率，远低于 20%，证明了神经网络模型选择阶段引入 G 分数的积极效果。因此，即使标签是不平衡的，在训练和验证阶段，适当的平衡技术和 G-Score 度量的组合使用已经导致了无偏的模型。此外，从实验结果来看，在每个模型中，沿着桶的精度值都呈现出积极的趋势，这表明我们的方法能够识别市场价格中最相关的变化。

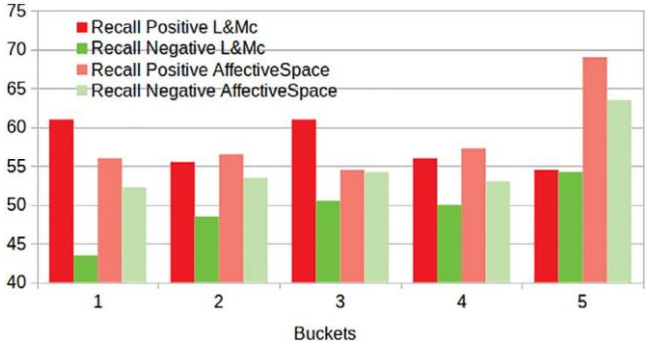
在评估了拉夫兰和麦当劳的字典功能后，对情感空间的使用进行了比较



图。5.沿着桶回忆:神经网络与 L&Mc。



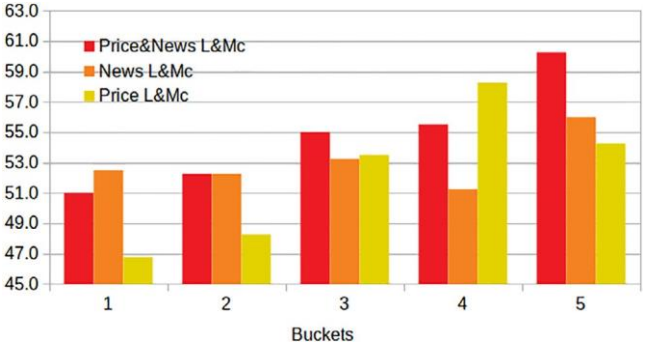
图。6.沿着桶的精度:L&Mc 与影响空间。



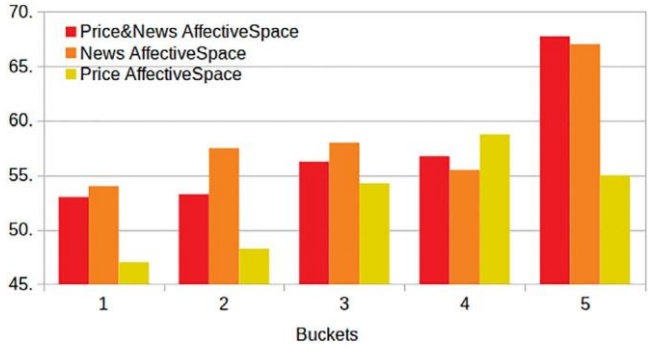
图。7.沿着桶回忆:L & Mc vs AffectiveSpace。

做过实验。除了神经元数量范围从[64，128，256]移动到[256，512，756]之外，神经网络使用相同的超参数集进行训练和验证。这是因为，根据 Panchal 和 Pan- chal (2014)给出的见解，神经网络中神经元的适当数量与输入向量的维数相关，在这种情况下，输入向量的维数从 151 增加到 611.分别地，维度 151 是用拉夫兰和麦当劳字典(111+40)获得的价格和新闻特征的连接，而 611 与来自影响-速度(111+500)的价格和新闻特征的连接相关。在训练和验证阶段之后，在样本外数据上测试模型。沿着桶的精确值和召回值，以及先前使用拉夫兰和麦克唐纳数据集获得的值，分别在图 1 和图 2 中报告。六和七。

该比较报告了在准确度值和正/负回忆增量方面的进一步改善。因此，这些特征
使用情感空间应用编程接口提取的信息更能有效地表达金融新闻文本中的情感。这



图。8.具有不同特性集的铲斗的精度。



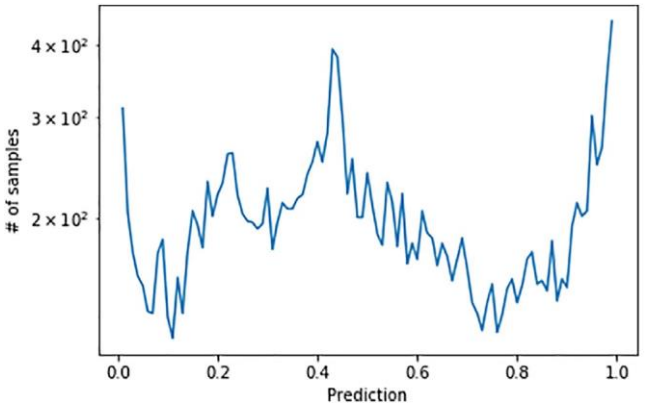
图。9.具有不同特征集的桶的精度影响空间。

是情感空间从结构化文本中提取概念层面情感的力量的证明。

在我们实验的第二阶段，研究了不同特征集之间的关系，以了解在市场趋势分类任务中包含情绪是否会导致更高的性能。为了实现这个目标，神经网络被训练和验证为与以前相同的超参数范围，但是基于从我们的数据中提取的三个不同的特征集，并且结果被分组在两个图表中。利用拉夫兰和麦克唐纳获得的性能和受影响的空间特征分别显示在图 1 和图 2 中。八和九。

通过对地块的仔细评估，发现《价格与新闻》集与《价格与新闻》集相比表现过度，但同时《价格与新闻》集与《新闻》集相比表现并不突出。这一行为表明，将情感表示添加到价格特征中会导致性能的重要提高。另一方面，这些结果表明，价格和新闻集之间的特征融合过程将允许模型更好地利用组合特征的代表性，而不是使用简单的串联。

在实验阶段的最后一部分，评估过程中的第二步，代表了我们方法的



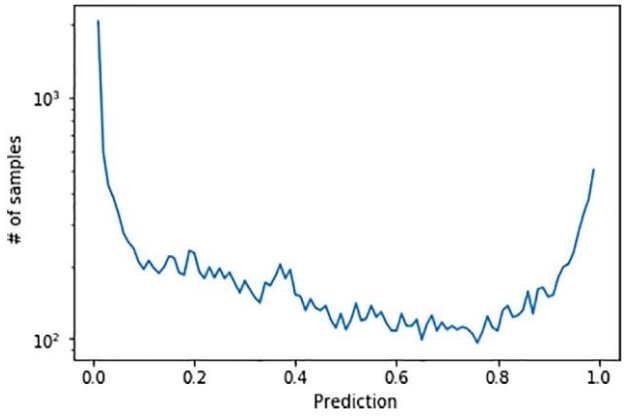


Table 4
Trading simulation results.

L&Mc		Annualized return	Maximum drawdown	Sharpe ratio
	Price	−46.5%	12.9%	-3.3%
	News	78.3%	**1.52%**	**7.78%**
	News&Price	**85.2%**	3.9%	4.76%
Affective		Annualized return	Maximum drawdown	Sharpe ratio
	Price	−45.6%	11,2%	−3.3%
	News	42.6%	3.82%	3.4%
	News&Price	5.05%	5.34%	0.3%
Buy & Hold		Annualized return	Maximum drawdown	Sharpe ratio
	Price	43.5%	1.59%	5.6%

一个新颖之处，是为了证明我们的预测在交易模拟中的作用。在这一阶段，按如下方式设置传统模拟。初始帐户值设置为 1，0000，000 美元，交易成本固定为 0.001 美元，这是从在线经纪人处检索到的。在计算夏普比率时，根据国债报价，无风险回报 r 的值被设定为 2.8%。如中所述，交易模拟是在计算了阈值总悬浮微粒后进行的

图。10.使用 L&Mc 数据集获得的预测:新闻(向上)和新闻&价格(向下)特征。

第三节。为了深入了解这一过程，图 2 显示了两个离散的概率分布。10 .

报告的图表还表明，利用价格和新闻结果的特征组合生成的预测分布更集中于输出层中 Sigmoid 激活函数的极值。这一事实突出表明，与仅使用新闻相比，使用价格和新闻设置训练的模型中的预测具有更高的可信度。

在阈值计算之后，利用样本外数据的预测来形成交易策略。如表 4 所示，我们的模型在应用于组合特征集时，实现了较高的正年化收益，证明了我们预测方法的有效性和强大的盈利能力。买入并持有策略是作为我们成就的财务基准而插入的。

综上所述，通过第二个评估步骤，我们的方法从开发的分类器的正确行为和在交易模拟期间获得的性能两方面被证明是有效的。此外，其他重要的成就，考虑到我们可用的功能集的预测能力，可以从报告的性能中得出。

在 Loughran 和 McDonald 字典数据集的情况下，利用组合特征集获得的增益最高，而使用 AffectiveSpace 2 时，新闻集的性能优于其他数据集。

这进一步支持了我们之前关于特征融合技术重要性的结论。事实上，当处理第一个数据集的 151 维输入时，我们的模型能够组合特征并获得更高的结果，即使特征集是简单连接的。

另一方面，当拉夫兰和麦克唐纳字典的 611 维向量作为输入时，我们的模型不能适当地利用这个更宽向量的特征组合能力。因此，在第二种情况下，特征融合过程将改善预测，从而带来更高的回报。其他业绩指标的数值得出了同样的结论。事实上，当使用价格和新闻数据集时，最大提款和夏普比率都揭示了投资组合价值的更多波动性。因此，即使特征的组合已经获得了非常高的回报，通过特征融合步骤可以提高我们感兴趣的成就，从而降低投资组合价值的波动性。

6. 结论和未来方向

这项工作的目标是通过使用应用于时间序列预测和情绪分析的机器学习技术，将技术和基本分析方法结合起来进行市场趋势预测。此外，我们旨在开发一个稳健的模型，能够预测股票投资组合的趋势，并在交易策略中利用其预测。因此，我们将前馈神经网络体系结构的开发提出为趋势分类问题，并分两步进行评估。执行第一步是为了评估分类器的统计行为，而第二步集中于测试模型预测在交易模拟中的有效性。

其结果是一个稳健的模型，能够有效和公平地对研究中的股票组合的正趋势和负趋势进行分类。通过评估正样本和负样本的召回值之间的差距来证明分类器的正确行为。此外，我们的模型能够识别市场趋势中最有意义的变化，并在交易模拟期间获得正回报。在这项工作中，使用了两种不同的方法从新闻中提取情感嵌入:拉夫兰和麦克登-阿尔德词典和情感空间 2.使用情感空间特征作为神经网络结构的输入，可以更有效地获得高精度值，而使用拉夫兰和麦克唐纳字典计算的特征的探索可以获得更高的年化收益值。对三个特征集(即价格、新闻和价格与新闻集)进行了比较，并将情感嵌入与价格技术指标相结合，结果仅超出了价格集的使用。另一方面，与单独使用新闻集相比，过度性能并不突出。因此，我们认为，为了解决这一弱点，使用适当的特征融合技术是一个有趣的未来研究方向。此外，我们意识到我们数据的时间跨度是有限的，我们正在积极努力检索更多的新闻数据，并在此基础上建立一个适当的回溯测试阶段来测试我们的模型。总之，这项工作为市场预测的技术和基础方法之间的未来合作奠定了坚实的基础。此外，它将分裂市场分析师半个多世纪的对立方法与利用机器学习和数据科学技术相调和。

竞争利益声明

我们声明没有利益冲突。

信用作者贡献声明

安德里亚·毕加索:软件，验证，形式分析，投资，资源，数据管理，写作-初稿，写作-评论和编辑，可视化。西蒙·梅雷洛:软件，验证，形式分析，调查，资源，数据管理，写作-初稿，写作-评论和编辑，可视化。马玉坤:监督，验证，形式分析，调查，写作-初稿。卢卡·奥内托:监督，方法学，形式分析，资源，写作-初稿，写作-评论和编辑。埃里克·坎布里亚:监督，方法，资源，写作-初稿，写作-评论和编辑。

参考

j .阿巴内尔。南。，&布什尔，b。J. (1998). 基本面分析策略的异常回报。*会计评论，19–45。*

Al Shalabi。，沙班，z。，& Kasasbeh，b。(2006). 数据挖掘:一个预处理引擎。*计算机科学杂志，2 (9)，735–739。*

阿明，a。，安华，s。，阿德南，a。，Nawaz，m。纽约州霍华德市。，Qadir，j . ，等。(2016). 处理类不平衡问题的对比过采样技术:客户流失预测案例研究。*IEEE 接入，4，7940–7957。*

班迪。(2007). *量化交易系统。*

j .博伦。，毛，h。，&曾，x。(2011). 推特情绪预测股市。*计算科学杂志，2 (1)，1–8。*

方框，g。，&詹金斯，g。(1970). 。

柬埔寨。，傅，j。，Bisio，f。，&茨苓，s。(2015). 情感空间 2:为概念层面的情感分析启用情感直觉.. *AAAI 人工智能会议。*

北查瓦拉。动词 （verb 的缩写）。，鲍耶，k。洛杉矶霍尔。，&凯格尔迈耶，w。页（page 的缩写）。(2002). Smote:合成少数过采样技术。*人工智能研究杂志，16，321–357。*

乔德里，r。，& Garg，k。(2008). 用于股票市场预测的混合机器学习系统。*世界科学、工程和技术学院，39 (3)，315–318。*

克里斯蒂安尼，纽约。，&肖-泰勒，j。(20 0 0). *介绍支持向量机和其他基于核的学习方法。*剑桥大学出版社。

西本科，g。(1989). 乙状线函数的叠加逼近。*控制、信号和系统数学，2 (4)，303–314。*

m .德乔杜里。，巽他拉姆，h。，约翰，a。，&塞利格曼，d。D. (2008). 博客交流动态能与股市活动相关吗？*超文本和超媒体会议。*

Fama，e . F. (1991). 有效资本市场:二。*金融杂志，46 (5)，1575-1617。*

Gentile。，&沃姆斯，m。K. (1999). 线性铰链损耗和平均余量。*神经信息处理系统的进展。*

哈扎里卡，d。，茨苓，s。，扎德，甲。坎布里亚，东。，Morency，l。页（page 的缩写）。，& Zimmer- mann，r。(2018). 对话视频中用于情感识别的会话记忆网络。在计算语言学协会北美分会的会议上:人类语言技术。

胡，z。，刘，w。，边，j。，刘，x。，&刘，t。(2018). 聆听混沌低语:面向新闻的股票趋势预测深度学习框架。*网络搜索和数据挖掘国际会议。*

黄，c。，杨，马超。，&创，y。(2008). 包装方法和复合分类器在股票趋势预测中的应用。*专家系统与应用，34 (4)，2870–2878。*

黄，w。，中森，y。，&王，s。(2005). 用支持向量机预测股市走势。*计算机与运筹学，32 (10)，2513–2522。*

赫斯特，h。E. (1951). 水库长期库容。*美国土木工程师学会会刊，116，770–799。*

Ioffe，s。&塞格迪，c。(2015). *批量标准化:通过减少内部协变量偏移来加速深度网络训练。*arXiv: 1502.03167。

金，f。，Self，n。，萨拉夫，p。巴特勒，p。，王，w。，&罗摩克里希南，n。(2013). 外汇电子出纳:使用新闻文章进行货币趋势建模。*知识发现和数据挖掘国际会议。*

Khaidem，l 。萨哈，美国，。& Dey，s。R. (2016). *用随机森林预测股票市场价格的走向。*arXiv: 1605.0 0 0 03。

金，金。，&韩，我。(20 0 0). 股价指数预测人工神经网络特征离散化的遗传算法方法。*专家系统与应用，19 (2)，125–132。*

金马，d。页（page 的缩写）,。&巴，j。(2014). *亚当:一种随机优化的方法。*arXiv: 1412 年.6980 .

m .库马尔。，& Thenmozhi，m。(2006). 股票指数走势预测:支持向量机和随机森林的比较。*工程研究与应用杂志。*

李，x。，谢，h。，陈，l。，王，j。，&邓，x。(2014). 基于情绪分析的新闻对股价回报的影响。*基于知识的系统，69，14–23。*

李，y。，潘，问。，杨，t。，王，s。，唐，j。，&坎布里亚，东。(2017). 学习用于情感分析的单词表示。*认知计算，9 (6)，843–851。*

Lo，a。W. (2004). 适应性市场假说:从进化的角度看市场效率。*投资组合管理杂志，即将出版。*

拉夫兰，t。，&麦当劳，b。(2011). 什么时候负债不是负债？文本分析、词典和 10-ks。*金融杂志，66 (1)，35–65。*

苏露莎，洛杉矶。，& r .布拉格斯。(2010). 高维类不平衡数据的类预测。*BMC 生物信息学，11 (1)，523。*

吕斯河。，& D'Aspremont，a。(2015). 利用文本分类预测新闻异常收益。*定量金融，15 (6)，999–1012。*

马，y。，彭，h。，&坎布里亚，东。(2018). 通过将常识知识嵌入到关注的 lstm 中进行有针对性的基于方面的情感分析。*AAAI 人工智能会议。*

Merello。，毕加索，a。，Oneto，l。，&坎布里亚，东。(2019). 预测未来市场趋势:哪个是最佳窗口？*国际神经网络学会大数据与深度学习。*

米塔尔，a。，& Goel，a。(2012). 利用推特情绪分析进行股票预测。*斯坦福大学，15 岁。*

水野彩香。，Kosaka，m。，Yajima，h。，& Komoda，n。(1998). 神经网络在股市预测技术分析中的应用。*信息与控制研究，7 (3)，111–120。*

Oancea，b。美国 Ciucu。C. (2014). *基于神经网络的时间序列预测。*arXiv: 1401.1333。

哦，c。，&盛，o。(2011). 探讨股票微博在预测未来股价走势中的预测力.. *国际信息系统会议。*

Panchal。，m & Panchal。(2014). 人工神经网络中隐节点数选择方法综述。*国际计算机科学和移动计算杂志，3 (11)，455–464。*

Patro，s。英国萨胡。K. (2015). *规范化:预处理阶段。*arXiv: 1503 年.06462 .

彭，y。&江，h。(2015). *利用单词嵌入和深度神经网络，利用金融新闻预测股价走势。*arXiv: 1506.07220。

毕加索，a。梅雷洛，s。，马，y。，马兰德里，l。，Oneto，l。，&坎布里亚，东。(2018). 市场趋势预测的技术分析和机器学习集成。*计算智能系列研讨会。*

茨苓，s。坎布里亚，东。，巴杰派，r。，&侯赛因，a。(2017). 情感计算述评:从单峰分析到多峰融合。*信息融合，37，98–125。*

钱，b。，&拉希德，k。(2004). 赫斯特指数与金融市场的可预测性。*金融工程与应用国际会议*。
饶，t。，& Srivastava，s。(2012). 用推特情绪分析来分析股市走势。*社会网络分析和挖掘进展国际会议*。
雷亚，r。(1993). *道氏理论:对其发展的解释，并试图定义其作为投机辅助工具的有用性*。弗雷泽出版公司。
洛克，d。W. 罗杰斯。K. ，卡巴冒险，m。，奥克斯利，m。E. ，& Suter，b。W. (1990). *多层感知器作为贝叶斯最佳鉴别函数的近似*。*IEEE 神经网络交易，1 (4)，296–298*。
鲁梅尔哈特博士。E. ，辛顿，g。E. ，&威廉姆斯。J. (1986). 通过反向传播错误学习表示。*自然，323 (6088)，533*。
舒马赫，r。，&陈，h。(2006). 利用财经新闻文章进行股市预测的文本分析。*美洲信息系统会议*。
舒马赫，r。页（page 的缩写）。，&陈，h。(2009). 基于突发金融新闻的股市预测文本分析:azfin 文本系统。*信息系统交易(TOIS)，27 (2)，12*。
希莫达拉，h。(20 0 0). 通过加权对数似然函数改进协变量移动下的预测推断。*统计规划和推断杂志，90 (2)，227–244*。
Si，j。，慕克吉，a。，刘，b。，李问。，李，h。，&邓，x。(2013). 利用基于话题的推特情绪进行股票预测。*计算语言学协会年会*。
北斯雷布罗。，雷尼，j。，& Jaakkola，t。南。(2005). 最大利润矩阵分解。*神经信息处理系统的进展*。
北斯里瓦斯塔瓦。，辛顿，g。，克里哲夫斯基，a。，苏斯克弗，我。，& r . Salakhutdinov。(2014). 辍学:防止神经网络过度拟合的简单方法。*机器学习研究杂志，15 (1)，1929-1958*。
泰，福。，&曹，l。(2001). 支持向量机在金融时间序列预测中的应用。*欧米茄，29 (4)，309–317*。
托马斯，m。(1999). 从业者方法和工具。*金融计算智能杂志，7 (3)，36–45*。
吴，j。长度。，苏，c。C. ，余，l。C. ，&常，p。C. (2012). 从股票新闻的情感分析和交易信息的技术分析利用组合特征预测股票价格。*经济学发展与研究国际会议录*。
邢，f。坎布里亚，东。维尔施公司。E. (2018). 基于自然语言的金融预测:综述。*人工智能评论，50 (1)，49–73*。
徐，y。，&科恩，s。B. (2018). 根据推文和历史价格预测股票走势。*计算语言学协会年会*。
姚，j。，& Poh，h。(1995). 用神经网络预测 klse 指数。*国际神经网络会议*。
姚，j。，谭，c。长度。，& Poh，h。(1999). 技术分析的神经网络:klci 研究。*国际理论和应用金融杂志，2 (02)，221–241*。
你，q。，罗，j。，金，h。，&杨，j。(2015). 使用渐进训练和域转移深度网络的鲁棒图像情感分析.. *AAAI 人工智能会议*。
张，g。页（page 的缩写）。(2003). 基于混合 arima 和神经网络模型的时间序列预测。*神经计算，50，159–175*。