

# 算法黑箱：网络传播误导性旧信息 操纵市场的研究<sup>\*</sup>

王 春

**内容提要：**智能互联网时代，新闻以机器学习为基础生产和推荐，公众难以了解其中环节，从而形成算法黑箱。算法黑箱以看似技术中立输出的与客观事实不符的信息，误导投资者决策，破坏市场定价功能，易引发系统性风险。本文以财经新闻文本相似性为网络传播误导性旧信息的代理变量，采用向量空间模型和词频-逆向文档频率法进行度量。基于数行者（深圳）科技有限公司开发的报刊新闻量化舆情数据平台中 1046620 条样本新闻数据，发现网络传播误导性旧信息会操纵资本市场，表现为：网络传播误导性旧信息发布一周后，公司股票超额收益反转，且财经新闻文本相似性越高，收益反转越明显；网络传播误导性旧信息还操纵了公司股票交易活动，即降低了公司股票收益波动性和超额交易量。将样本公司按机构投资者持股比例分组后，发现机构投资者持股比例低的公司受网络传播误导性旧信息操纵的影响更大。网络传播误导性旧信息操纵市场的机制在于个人投资者行为偏见，即个人投资者混淆了财经新闻中包含的新信息与旧信息，并且按照旧信息进行交易。本文为互联网平台算法生产的财经新闻纳入监管提供了政策依据。

**关键词：**网络信息 文本分析法 财经新闻 市场操纵

**作者简介：**王 春，南京审计大学金融学院副教授、硕士生导师，211815。

**中图分类号：**F830.9 **文献标识码：**A **文章编号：**1002-8102(2021)11-0085-16

## 一、引 言

从传感器新闻、大数据分析到算法推荐，人工智能正带来一场媒体技术革命。它极大地提高了信息的生产和传播效率，但也引发了监管部门和用户对这一新技术应用中出现的算法黑箱问题的关切和疑虑。例如，智能互联网通常以机器学习或神经网络为基础来生产和推荐算法新闻。由于算法技术本身的复杂性，受众很难像了解传统媒体运作流程那样去了解算法新闻的生产过程，

<sup>\*</sup> 基金项目：江苏省社会科学基金项目“市场操纵视角下网络传播误导性信息识别及风险防范研究”（19EYD004）；江苏高校优势学科建设工程三期项目“南京审计大学应用经济学”（苏政办发〔2018〕87号）。感谢匿名审稿人的宝贵意见，文责自负。王春电子邮箱：wangchun0791@163.com。

因此这个中间环节就被称为“算法黑箱”。<sup>①</sup>

算法黑箱的主要风险在于设计者用看似客观的技术中立算法代替新闻的各个环节,输出与客观事实不符的结果,具有极大的欺骗性。英国文化研究专家 Lash(2007)强调,“在一个媒体和代码无处不在的社会,权力存在于算法之中”。算法自身的高门槛、神秘性为其披上了“黑箱”的外衣,其标榜的技术客观性会对客观真实造成真正的伤害。例如,2017年7月5日下午4时许,一篇名为《传复星集团董事长失联,交易所11复星债与10复星债大跌》的文章在财经媒体迅速传播。受此消息影响,7月6日,相关概念A股及港股盘中股价大跌,对相关板块指数造成明显影响,多只跟踪指数ETF基金回撤明显。证监会随即启动立案,开展信息溯源和核查工作。稽查部门调查发现,7月5日,同花顺网络运营的同花顺财经网站使用网络爬虫软件,从“和讯债券”网站自动抓取了一条2015年12月发布的关于复星集团董事长郭广昌失联的陈旧信息。该网站工作人员在文章录入审核中未发现信息来源陈旧、时间不匹配等问题,将早已过期的文章作为即时新闻在同花顺财经网站债券栏目同步发布,导致与当下现实情况明显不符的信息迅速传播,严重误导市场投资者,扰乱了资本市场秩序,造成了恶劣影响。<sup>②</sup>

如前文所述,财经网站平台因算法黑箱过程,推出了干扰客观事实、误导市场其他投资者的算法新闻,从而形成了“市场操纵”。在经济金融方面,市场操纵主要是指市场上所使用的不恰当方式,可分为多种类型<sup>③</sup>(Putnins, 2012)。本文关注的是信息型市场操纵。

党的十九届五中全会关于“十四五”规划建议明确提出,要健全金融风险预防、预警、处置体系,对违法违规行为零容忍。构建金融有效支持实体经济的体制机制,提升金融科技水平。随着智能互联网时代的到来,运用机器学习、数据挖掘等现代科技手段识别算法黑箱生产的误导性旧信息、实时监测资本市场交易过程、及时发现涉嫌市场操纵的违法违规行为,能够有效提升金融风险防范能力,在新发展格局中支持经济高质量发展,同时也有助于提升监管工作的数字化、智能化水平,促进监管科技创新。

本文研究发现:(1)网络传播误导性旧信息会操纵公司股票收益,在未来一周,公司股票超额收益下降,且财经新闻文本相似性越高,收益反转越明显;(2)网络传播误导性旧信息会操纵公司股票交易活动,即降低公司股票收益波动性和超额交易量;(3)机构投资者持股比例较低的公司受网络传播误导性旧信息操纵的影响更大;(4)网络传播误导性旧信息操纵市场的行为机制是个人投资者混淆了财经新闻中包含的新信息与旧信息,并且按照旧信息进行交易。也正是这种认知偏差,使得网络传播误导性旧信息操纵股价现象得以发生。

本文的研究贡献在于以下几个方面。(1)市场操纵识别方法方面。将财经新闻文本相似性作为算法黑箱产生的误导性旧信息操纵股价的识别指标。采用向量空间模型(VSM)来表示新闻的文字内容,并利用余弦相似性来衡量两篇新闻的内容相似程度,与Tetlock(2011)的方法不同,本文还使用词频-逆向文档频率(TF-IDF)法赋予词组在不同文档向量中的权重,使得财经新闻文本相似性的度量更加准确。与现有文献认为的媒体情绪操纵(罗伯特·J.希勒,2007)、公司财务操纵(Huang等,2014)、收盘价操纵(李志辉等,2018)不同,本文认为文本相似性高的旧财经新闻是误导性信息,会干扰投资者,因而是市场操纵的识别指标。(2)市场操纵理论机制方面。传统的

① 黄晓勇:《认知智能媒介的阶段性风险》,中国社会科学网,2019年9月19日。

② 《证监会查获一起网络传播媒介传播误导性信息案件》,中国证监会网站,2017年7月21日。

③ 具体分为拉升出货型(Pump and Dump)、合约型(Contract-based)和市场力量型(Market Power)。其中,拉升出货型又包含交易型(Trade-based)、信息型(Info-based)和行动型(Action-based)。

操纵在于利用虚假信息有意做空或拉高股价,从而使自己获利(Vila, 1989; Bagnoli 和 Lipman, 1996),或者公司经理与经纪商共谋操纵股价以获利(Khwaja 和 Mian, 2005)。本文认为在智能互联网时代,算法黑箱过程利用爬虫技术将过去的新闻信息抓取出来,再次发布在网络上,也会实现股价操纵。这是由于个人投资者行为偏见产生作用,即个人投资者无法辨别新信息与旧信息,其后果表现为降低公司股票收益波动性和超额交易量。(3)媒体信息定价方面。与 Birz(2017)、游家兴和吴静(2012)、余峰燕等(2012)主要采用公司全称或公司股票简称作为关键词搜索方式获取样本不同,本文以互联网信息为对象,采用以余弦相似性和词频-逆向文档频率法为基础产生的大数据,相较于以往研究范式中数据的收集、分析与处理方法能够产生积极影响,推动资本市场网络信息操纵研究不断发展。

本文剩余内容安排如下:第二部分为文献回顾,第三部分为理论模型与研究假说,第四部分为研究设计,第五部分为实证分析,第六部分为网络传播误导性旧信息操纵市场的机制分析,第七部分为结论与启示。

## 二、文献回顾

### (一)市场操纵识别方法

市场操纵识别方法依照识别客体的不同,主要分为模型模拟识别方法、机器学习识别方法以及诉讼案件识别方法等。(1)模型模拟识别方法。Fruth 等(2013)利用模型模拟的方法来识别市场操纵,提出一个允许时间依赖、价格深度确定、订单具有充分弹性的有限订单预订模型,在没有订单预订价差的情况下,如果依赖模型参数进行交易,就会产生价格操纵。(2)机器学习识别方法。Li 等(2017)基于中国证监会发布的信息和数据,使用机器学习方法识别市场操纵。(3)诉讼案件识别方法。Neupane 等(2017)、Comerton-Forde 和 Putniņš(2011)通过大量股票 IPO、尾市交易类市场操纵诉讼案件,提取相应的交易量特征,以识别未来交易中可能出现的市场操纵。

### (二)市场操纵理论机制

在资本市场领域,市场操纵问题一直存在。当前围绕市场操纵理论机制的研究文献包括以下几个方面。(1)管理层操纵机制。Peng 和 Roell(2013)、Schroth(2018)针对管理层短期股价操纵与操纵倾向不确定性现象,提出了最优补偿理性预期模型,比较了按长期业绩与短期业绩提供报酬的工资合约、市场操纵以及投资者不确定性对股价信息含量的影响。(2)盈余操纵机制。Wu 等(2012)、Strobl(2013)认为,股票期权作为补偿机制在公司管理中被广泛使用,它将管理层财富与股东财富联系在一起,其内在价值取决于股票价格与行权价格的差值。高管作为自利的个体,会操纵公司盈余报告以影响股价。(3)股票拆分机制。Kim 等(2012)认为,股票拆分之后,有后续新闻报道的公司存在利用股票拆分手段操纵市场的可能。(4)共同基金操纵机制。Ben-David 等(2013)认为,有些共同基金在关键的披露期会操纵股票价格。

### (三)陈旧新闻对股市的反应

上市公司相关信息流随着时间的推移逐渐被投资者接受,操纵者所拥有的不为大众所知的信息也会随着信息的扩散变为旧信息,但还具有一定的市场影响(Bernhardt 和 Miao, 2005)。Tetlock(2011)首次以文本相似性区分新旧信息,并考察了旧信息对上市公司的影响。Loughran 和 McDonald(2011)以上市公司年报文本相似性考察了管理层责任缺失下的市场反应。余峰燕等(2012)考察了大众媒体重复旧信息行为对资产价格的影响。Birz(2017)按新闻报纸及网络标题分

类,考察了旧新闻对资产价格收益的影响。

#### (四)对现有研究的评述

当前市场操纵主要从识别方法、理论机制与旧新闻的市场反应方面进行了一系列研究,这在传统媒体信息发布时代起到了一定的积极作用。然而,随着智能互联网时代的到来,出现了算法新闻和算法黑箱,这需要用大数据技术采集网络新闻,结合机器学习和自然语言处理技术,构建向量空间模型予以识别。在理论机制方面,结合算法黑箱过程产生误导性信息操纵市场的案例情况,需要从新闻文本相似性导致投资者混淆新旧信息,从而按照旧信息进行交易来进行刻画。在旧新闻的市场反应方面,传统的按照新闻标题区分新旧新闻的方法,准确性下降且样本规模有限,在大数据背景下,数据获得的速度、容量和质量都有所提升,新旧网络新闻从文本方面予以识别,更加符合智能媒介传播的特征。

### 三、理论模型与研究假说

在过度反应和不足反应(Barberis等,1998;Daniel等,1998;Brunnermeier,2005)以及投资者异质性(Hong和Stein,1999)的基础上,假定:(1)市场上只存在机构投资者和个人投资者两类投资者,个人投资者有限理性程度为 $m$ ,机构投资者有限理性程度为 $1-m$ ,两类投资者都满足常数绝对风险厌恶(CARA)性质;(2)所有投资者都是价格接受者;(3)单一资产股利 $d$ 服从正态分布,且股利分三期向投资者发布。

依上述假定,投资者将在第1期观察到信号 $s_1$ ,第2期观察到信号 $s_1+s_2$ ,第3期观察到信号 $s_1+s_2+s_3$ 。两类投资者仅在包含网络传播误导性信息(下称“旧信息”)  $s_1$  和新信息  $s_2$  的第2期,处理接收到的信号的方式不同。机构投资者能够完全分清这两类信息,忽略旧信息  $s_1$ ,接受新信息  $s_2$ ;而个人投资者却接受新信息  $ks_1+s_2$ ,其中  $0 < k < 1$ ,他们会对部分旧信息做出反应。参数  $k$  表示个人投资者对旧信息反应的程度,当  $k=0$  时,意味着该投资者属于机构投资者。

两类投资者都在各期依据可得信息,考虑资产需求以最大化其CARA效用函数。为精简模型均衡解,只考虑投资者最大化下一期期望财富。为消除风险厌恶对资产价格的影响,同时假定:机构投资者和个人投资者的风险厌恶参数相同;单一资产在净供给为零时也可以获得。使用传统的后向归纳法,得到两期市场均衡的出清价格:

$$p_2 = (1 + m \times k) s_1 + s_2 \quad (1)$$

$$p_1 = [1 + m \times (1 - m) \times k] \times s_1 \quad (2)$$

在第0期,所有投资者拥有相同的信息,价格为0。当 $m=0$ 时,股票市场上只有机构投资者,机构投资者遵循贝叶斯准则进行决策,且具备无限认知能力,方程(1)变为 $p_2 = s_1 + s_2$ ,方程(2)变为 $p_1 = s_1$ 。当 $m>0$ 时,相对于 $m=0$ ,第1期价格对信号 $s_1$ 产生过度反应。在第2期,这种过度反应继续存在,并且因个人投资者的参与而增强。个人投资者在第1期和第2期产生过度反应的根本原因在于未能辨认旧信息和新信息。

大量经验事实表明,两则新旧新闻事件信号 $s_1$ 和 $s_1+s_2$ 之间的收益协方差 $Cov(d-p_2, p_2-p_1)$ 、 $Cov(d-p_2, p_1-p_0)$ 为负,且呈负相关关系。<sup>①</sup> 根据方程(1)和方程(2),可以得到信号 $s_1$ 和信号 $s_1+$

<sup>①</sup> 一般认为,收益反转意味着过度反应。

$s_2$  之后的误导性新闻事件收益反转表达式:

$$Cov(d - p_2, p_2 - p_1) = -m^2 k^2 \sigma_1^2 \leq 0 \quad (3)$$

$$Cov(d - p_2, p_1 - p_0) = -mk[1 + m(1 - m)k] \sigma_1^2 \leq 0 \quad (4)$$

方程(3)和方程(4)表明,在第2期旧信息发布后,无论是否预料到,对旧信息的过度反应都导致收益反转。可以看到,首先,当有大量的信息重新发布,尤其是当这些旧信息看起来像新信息时,即  $\sigma_1^2$  越大,收益反转也越大。其次,在第二次释放信号  $s_1$  时,参与股票市场的个人投资者越多,即  $m$  越大,收益反转也越大。由此,本文提出假设1。

假设1:网络传播误导性旧信息会操纵公司股票收益,在未来一周,公司股票超额收益将反转,且误导性信息文本相似性越高,收益反转越明显。

Bernhardt 和 Miao(2005)认为当信息分阶段向投资者披露时,如果信息是异质性的,比如临近收益公告的时间点,往往内容有所不同,会提升交易量和股票收益波动性。而当内容相似,即信息同质时,股票收益波动性会下降。Gropp 和 Kadareja(2012)将信息分为私人信息和公共信息,其中公共信息是可观测的,投资者普遍知道;私人信息则是不可观测的,并非所有投资者都知道。伴随着公共信息的扩散,股票收益波动性提高;而当公共信息为众多投资者所知晓时,股票收益波动性则开始下降。当前网络传播信息明显提升了股票市场定价效率(丁慧等,2018)。因此,新消息首次在网络公布时,很快反映在股价中,当网络抓取新闻文本再次发布时,旧消息使得股票收益波动性降低。此外,误导性信息文本相似也使得投资者关注度下降,收益率降低。由此,本文提出假设2。

假设2:网络传播误导性旧信息操纵了公司股票交易活动,即降低了公司股票收益波动性和超额交易量。

## 四、研究设计

### (一)数据来源与样本选取

本文数据来自数行者(深圳)科技有限公司开发的报刊新闻量化舆情数据平台。该平台由香港中文大学利用大数据技术采集并分析主要新闻媒体<sup>①</sup>网站上发布的与A股全部上市公司相关的新闻,结合机器学习和自然语言处理技术对收集到的上千万篇原始新闻进行严谨的清理分析及核验,有效地解决了新闻与目标公司的匹配问题。本文数据在实际处理过程中,通过新闻标识(*News\_id*)变量将报刊新闻量化舆情数据对应匹配,然后通过新闻发布日期(*Pub\_Date*)和上市公司代码(*Stkcd*)变量将处理过的报刊新闻量化舆情数据与股票市场交易数据匹配,股票市场交易数据来源于RESSET股票数据库。样本包含2017年1月1日至2017年12月31日时间范围的1046620条新闻数据。样本数据处理与多元回归分析均使用SAS 9.4软件。为控制异常值影响,本文对所有连续变量采取1%和99%缩尾(Winsorize)处理。此外,剔除了交易数据缺失值和金融行业上市公司数据。

### (二)变量定义

#### 1. 原始收益率

原始收益率( $Ret[2,5]$ )是指网络传播误导性旧信息发布后第2天到第5天的收益率,排除 $t +$

① 主要新闻媒体有《21世纪经济报道》《上海证券报》《上海金融报》《中国证券报》《证券日报》《证券时报》等。

1, 因为临近事件日买卖差价的反弹会减少股票收益。

## 2. 公司股票交易活动

公司股票交易活动包含收益波动性和超额交易量两个指标。

(1) 收益波动性。收益波动性 ( $|AbRet_{it}|$ ) 即公司  $i$  在  $t$  日的绝对异常收益率, 等于公司原始收益率减去相应的上证综合指数或深圳成分指数收益率的绝对值。考虑到股票市场调整已经反映了大部分公司的系统性波动, 本文不再使用 Fama-French 三因子、四因子甚至更复杂的五因子模型进行波动性调整, 因为来自公司层面的特定新闻事件引起的收益波动不能被传统风险因子所解释 (Tetlock 等, 2008)。而且后文的检验将公司规模列为自变量, 这已经反映了来自市场贝塔的期望收益升水 (Fama 和 French, 1992)。

(2) 超额交易量。超额交易量 ( $AbTurn_{it}$ ) 由异常换手率度量, 等于公司  $i$  在  $t$  日换手率的对数减去  $[t-5, t-1]$  日内平均换手率的对数。

## 3. 网络传播误导性旧信息

本文研究的网络传播误导性旧信息是指由网络传播文本相似性信息带来的误导投资者的旧新闻。网络传播误导性旧信息由财经新闻文本相似性 (*Similarity*) 指标度量。<sup>①</sup> 采用向量空间模型来表示新闻的文字内容, 并利用余弦相似性来衡量两篇新闻的内容相似程度。特别地, 在对一篇新闻正文完成分词后, 这篇新闻被分割为若干词语的序列。每一篇新闻文档  $d_j$  可根据该序列中的词语表示为一个如下的向量, 该向量中的每一维都对应一个特定的词语权重。

$$d_j = (w_{1j}, w_{2j}, \dots, w_{ij}) \quad (5)$$

此外, 本文采用词频-逆向文档频率法赋予一个词组在不同文档向量中的权重。具体地, 一个词语  $t$  在文档  $d$  中的权重计算公式为:

$$w_{t,d} = f_{t,d} \cdot \log \frac{|D|}{|\{d' \in D \mid t \in d'\}|} \quad (6)$$

其中,  $f_{t,d}$  是词组  $t$  在文档  $d$  中出现的频率 (局部参数),  $\log \frac{|D|}{|\{d' \in D \mid t \in d'\}|}$  是逆向文档频率 (全局参数)。 $|D|$  是文档集中的文档总数,  $|\{d' \in D \mid t \in d'\}|$  是含有词组  $t$  的文档数。对于给定任意两篇新闻的文档向量  $d_k$  和  $d_j$ , 其余弦相似性为:

$$Similarity[d_j, d_k] = \frac{d_j \cdot d_k}{\|d_j\|_2 \|d_k\|_2} \quad (7)$$

其中, 两篇文章的余弦相似性分数取值范围在  $-1$  与  $1$  之间, 分数越接近  $1$ , 则表示两篇文章越相似。

## 4. 超额收益率

超额收益率 ( $AbRet_{it}$ ) 为公司  $i$  在  $t$  日的原始收益率减去相应的上证综合指数或深圳成分指数

<sup>①</sup> Tetlock (2011)、余峰燕等 (2012) 通过当前与先前 (至少一个交易日前) 对同一公司信息表达的文本相似性来界定新闻信息的新与旧。本文沿用他们的方法, 构建 *Similarity* 指标, 界定新信息与旧信息。*Similarity* 越高, 表示新闻内容越陈旧; *Similarity* 越低, 表示新闻内容越新。此外, 对新闻数据按以下原则进行了清理: (1) 将广告、公司公告、市场行情汇报等非相关新闻予以剔除; (2) 将正文少于 20 个字的新闻予以剔除; (3) 将没有真正提及任何一家上市公司的新闻予以剔除; (4) 将历史新闻相似度低于 0.4 的数据予以剔除; (5) 将文本中含有“转载、转引及新闻总结”类词语文本的财经新闻予以剔除。



收益率。

#### 5. 新闻文本情绪

本文采用有监督的机器学习模型,以及香港中文大学授权使用的数据集,该数据集经过多位经管类专业学生的交叉标注验证,准确度高。另外,我们还从证券时报官网等权威信息源收集了一些公开的标注数据集,从而使整个训练数据集包括 10 余万条准确的训练样本,准确率在 90% 以上。利用机器学习技术中的支持向量机方法,将一个句子所表达的情感倾向性分为三类:正面、中性和负面。然后,根据句子中所有词的词向量(Word Vector)统计值,经过式(8)调整,得到标准化的网络误导性信息文本情绪,即新闻文本情绪(Sentiment)。

$$Sentiment = \ln(1 + \frac{\text{正面句子数} - \text{负面句子数}}{\text{正面句子数} + \text{负面句子数} + 1}) \quad (8)$$

#### 6. 控制变量

控制变量包括  $t$  日新闻提及公司次数( $Ment\_Times_{it}$ )、事件日前一周新闻提及公司异常次数( $Ment\_Times_{it}[-5, -1]$ )。按 Tetlock (2007) 的界定,控制变量还包括  $t$  日新闻文本情绪( $Sentiment_{it}$ )。此外,为了控制 Tetlock (2007) 界定的新闻文本情绪的影响,本文的控制变量还包括  $t$  日超额收益率与新闻文本情绪的交互项( $AbRet_{it} \times Sentiment_{it}$ )、 $t$  日超额收益率与  $t$  日新闻提及公司次数的交互项( $AbRet_{it} \times Ment\_Times_{it}$ )。

#### 7. 市场控制变量

市场控制变量包括  $t-1$  日公司规模( $MktCap_{i,t-1}$ )、 $[t-5, t-1]$  日累计市场调整收益率( $AbRet_{it}[-5, -1]$ )、 $[t-5, t-1]$  日平均市场调整波动率( $Volat_{it}[-5, -1]$ )、 $[t-5, t-1]$  日 Amihud's 非流动性指标( $Illiq_{it}[-5, -1]$ )(孔东民等,2019)以及超额交易量( $AbTurn_{it}$ )。考虑到大市值公司与小市值公司反转的可能性不同,将公司  $i$  在  $t$  日收益波动性与公司规模 of 的交互项( $|AbRet_{it}| \times MktCap_{i,t-1}$ )也纳入其中。变量定义和描述见表 1。

表 1 变量定义和描述

| 变量名称       | 变量符号                  | 变量定义   |
|------------|-----------------------|--|
| 被解释变量      |                       |  |
| 原始收益率      | $Ret$                 | 网络文本相似新闻发布后,样本期间该公司原始收益率                           |
| 超额交易量      | $AbTurn$              | 上市公司在事件日换手率的对数减去前一周平均换手率的对数                        |
| 收益波动性      | $ AbRet $             | 公司原始收益率减去相应指数收益率的绝对值                               |
| 解释变量       |                       |  |
| 财经新闻文本相似性  | $Similarity$          | 任意两篇新闻文档向量的余弦相似性                                   |
| 控制变量       |                       |  |
| 新闻提及公司次数   | $Ment\_Times$         | 新闻提及该上市公司次数的对数                                     |
| 新闻提及公司异常次数 | $Ment\_Times[-5, -1]$ | 一周前平均每日公司网络新闻条数减去 $[t-30, t-6]$ 时间窗口内每日公司网络新闻条数的对数 |
| 新闻文本情绪     | $Sentiment$           | 新闻文本句子所表达的情感倾向性                                    |

续表 1

| 变量名称            | 变量符号                  | 变量定义                           |
|-----------------|-----------------------|--------------------------------|
| 市场控制变量          |                       |                                |
| 公司规模            | <i>MktCap</i>         | 上市公司市场价值的对数                    |
| 累计市场调整收益率       | <i>AbRet</i> [-5, -1] | 事件日前一周,公司累计超额收益率               |
| 市场调整波动率         | <i>Volat</i> [-5, -1] | 事件日前一周,股票日波动率减去相应指数波动率         |
| Amihud's 非流动性指标 | <i>Illiq</i> [-5, -1] | 事件日前一周,当日该公司股票交易额除以当日收益率对数的绝对值 |

## 五、实证分析

### (一)描述性统计

表 2 报告了财经新闻文本相似性(*Similarity*)、原始收益率(*Ret*)、超额交易量(*AbTurn*)以及收益波动性( $|AbRet|$ )等变量的描述性统计,包括各变量的均值、标准差、最小值及最大值。由表 2 可知,*Similarity* 的有效样本为 1046620 条新闻,最大值为 0.98,说明有高度相同的旧信息作为新闻重新发布。 $[t-5, t-1]$  日累计市场调整收益率(*AbRet*[-5, -1])的均值为 0.006,  $[t+2, t+5]$  日原始收益率(*Ret*)的均值为 -0.001,说明文本相似的旧信息发布一周前公司股票收益可能上升,一周后收益有反转下降的可能。

表 2 变量总体描述性统计

| 变量                    | 观察数     | 均值     | 标准差    | 最小值     | 最大值     |
|-----------------------|---------|--------|--------|---------|---------|
| <i>Ret</i>            | 922009  | -0.001 | 0.023  | -0.185  | 0.151   |
| <i>AbTurn</i>         | 912575  | 0.181  | 2.597  | -42.724 | 81.693  |
| $ AbRet $             | 911790  | 0.031  | 0.044  | 0       | 0.467   |
| <i>Similarity</i>     | 1046620 | 0.536  | 0.132  | 0.4     | 0.98    |
| <i>Ment_Times</i>     | 1193582 | 4.856  | 10.519 | 0       | 212     |
| <i>Sentiment</i>      | 1193582 | 0.319  | 0.555  | -0.989  | 0.994   |
| <i>MktCap</i>         | 1079057 | 24.287 | 1.858  | 16.684  | 28.1671 |
| <i>AbRet</i> [-5, -1] | 891871  | 0.006  | 0.064  | -0.41   | 0.692   |
| <i>Volat</i>          | 907883  | 0.018  | 0.01   | 0       | 0.087   |
| <i>Illiq</i>          | 912552  | 24.354 | 2.2    | 11.595  | 44.129  |

注:由于停牌和剔除金融业等原因,会显示缺失值,因而观察数会不一样。

表 3 报告了各主要变量的 Pearson 相关系数分析。从表 3 可以看到三个特征:第一,财经新闻文本相似性(*Similarity*)与  $[t+2, t+5]$  日的原始收益率(*Ret*)负相关,系数为 -0.004;第二,财经新闻文本相似性(*Similarity*)与超额交易量(*AbTurn*)负相关,系数为 -0.015;第三,财经新闻文本相似性(*Similarity*)与收益波动性( $|AbRet|$ )负相关,系数为 -0.006。这说明财经新闻文本相似性可能操纵公司股票收益,在未来一周,公司股票超额收益下降。此外,财经新闻文本相似性还可能操纵相应的公司股票交易活动,即降低公司股票收益波动性与超额交易量。



表 3 主要变量 Pearson 相关系数分析

| 变量                      | <i>Ret</i> | <i>AbTurn</i> | $ AbRet $  | <i>Similarity</i> | <i>Ment_</i><br><i>Times</i> | <i>Sentiment</i> | <i>MktCap</i> | <i>AbRet</i><br>[ -5, -1 ] | <i>Volat</i> |
|-------------------------|------------|---------------|------------|-------------------|------------------------------|------------------|---------------|----------------------------|--------------|
| <i>AbTurn</i>           | -0.051 *** |               |            |                   |                              |                  |               |                            |              |
| $ AbRet $               | 0.241 ***  | 0.118 ***     |            |                   |                              |                  |               |                            |              |
| <i>Similarity</i>       | -0.004 *** | -0.015 ***    | -0.006 *** |                   |                              |                  |               |                            |              |
| <i>Ment_Times</i>       | 0.007 **   | 0.012 **      | 0.052 **   | 0.009 *           |                              |                  |               |                            |              |
| <i>Sentiment</i>        | 0.016 ***  | -0.013 **     | 0.006 *    | 0.057 **          | -0.159 **                    |                  |               |                            |              |
| <i>MktCap</i>           | 0.025 **   | -0.09 ***     | -0.211 **  | -0.012 *          | 0.083 **                     | 0.071 **         |               |                            |              |
| <i>AbRet</i> [ -5, -1 ] | 0.018 ***  | 0.321 ***     | 0.261 ***  | -0.026 ***        | 0.015 ***                    | 0.042 ***        | -0.016 ***    |                            |              |
| <i>Volat</i>            | -0.037 *** | 0.036 ***     | 0.312 ***  | -0.034 ***        | 0.036 ***                    | -0.053 ***       | -0.395 ***    | 0.097 ***                  |              |
| <i>Illiq</i>            | 0.008 **   | -0.093 ***    | -0.157 *** | -0.021 **         | 0.081 ***                    | 0.053 ***        | 0.778 ***     | -0.072 **                  | -0.234 ***   |

注：\*\*\*、\*\*和\*分别表示在1%、5%和10%的水平下显著。下同。

## (二)模型设定

### 1. 网络传播误导性旧信息操纵公司股票收益分析

为了验证假设1,即网络传播误导性旧信息通过财经新闻文本相似性干扰投资者交易决策,进而操纵公司股价,本文主要考虑一周内的收益反转情况,参考 Fama 和 MacBeth(1973)的日截面回归模型,建立如下模型:

$$\begin{aligned}
 Ret[2,5] = & \alpha + \beta_1 \times AbRet_{i,t} + \beta_2 \times AbRet_{i,t} \times Similarity_{j,k} + \beta_3 \times Similarity_{j,k} + \\
 & \beta_4 \times AbRet_{i,t} \times Ment\_Times_{i,t} + \beta_5 \times Ment\_Times_{i,t} + \beta_6 \times AbRet_{i,t} \times \\
 & Sentiment_{i,t} + \beta_7 \times Sentiment_{i,t} + \beta_8 \times AbRet_{i,t} \times MktCap_{i,t-1} + \\
 & \beta_9 \times MktCap_{i,t-1} + \beta_{10} \times AbRet_i[-5, -1] + \beta_{11} \times AbTurn_{i,t} + \\
 & \beta_{12} \times Volat_i[-5, -1] + \beta_{13} \times Illiq_i[-5, -1] + Ind_{i,t} + \varepsilon_{i,t-1,t,j,k}
 \end{aligned} \quad (9)$$

其中, $Ind_{i,t}$ 为行业控制变量。这里主要考察式(9)中的两个自变量: $t$ 日公司*i*的超额收益率( $AbRet_{i,t}$ )、 $t$ 日公司*i*的超额收益率与公司*i*财经新闻文本相似性的交互项( $AbRet_{i,t} \times Similarity_{j,k}$ )。如果系数 $\beta_1$ 为负,则表示未来一周( $[t+2, t+5]$ )收益反转;如果系数 $\beta_1$ 为正,则表示未来一周收益状态持续。如果 $\beta_2$ 显著为负,说明财经新闻文本相似性越高,收益反转越明显;而如果 $\beta_2$ 显著为正,则说明财经新闻文本相似性越高,收益反转情况越差。

### 2. 网络传播误导性旧信息操纵公司股票交易活动分析

为了验证假设2,即网络传播误导性旧信息操纵公司股票交易活动的情况,同样参考 Fama 和 MacBeth(1973)的日截面回归模型,建立如下模型:

$$\begin{aligned}
 |AbRet_{i,t}| = & \alpha + \beta_1 \times Similarity_{j,k} + \beta_2 \times MktCap_{i,t-1} + \beta_3 \times AbRet_i[-5, -1] + \beta_4 \times \\
 & Volat_i[-5, -1] + \beta_5 \times Illiq_i[-5, -1] + \beta_6 \times Ment\_Times_{i,t} + \beta_7 \times \\
 & Sentiment_{i,t} + Ind_{i,t} + \varepsilon_{i,t-1,t,j,k}
 \end{aligned} \quad (10)$$

$$\begin{aligned}
 AbTurn_{i,t} = & \alpha + \beta_1 \times Similarity_{j,k} + \beta_2 \times MktCap_{i,t-1} + \beta_3 \times AbRet_i[-5, -1] + \beta_4 \times \\
 & Volat_i[-5, -1] + \beta_5 \times Illiq_i[-5, -1] + \beta_6 \times Ment\_Times_{i,t} + \\
 & \beta_7 \times Sentiment_{i,t} + Ind_{i,t} + \varepsilon_{i,t-1,t,j,k}
 \end{aligned} \quad (11)$$

这里主要考察式(10)与式(11)中的系数 $\beta_1$ ,如果系数 $\beta_1$ 为负,说明网络传播误导性旧信息操纵了公司股票交易活动,即降低了公司股票收益波动性与超额交易量;如果系数 $\beta_1$ 为正,说明网络传播误导性旧信息促进了公司股票交易活动,即提高了公司股票收益波动性与超额交易量。

### (三)回归分析

根据上述理论模型的假定,个人投资者和机构投资者对新旧信息的反应不同,且考虑到在互联网时代,搜索引擎的出现为投资者收集和获取信息提供了便利,我国投资者具有较强的信息获取能力(冯旭南,2014),当出现的信息对其投资行为影响较严重时,投资者可能会通过网络搜索来辨别信息的真伪。因此,本文将样本公司按报道时季度机构投资者持股比例分为高机构投资者组和低机构投资者组,并观测低机构投资者组是否受误导性信息操纵的影响更大。

#### 1. 网络传播误导性旧信息操纵公司股票收益分析

表4报告了对假设1的检验结果。

表4 网络传播误导性旧信息与公司股价

| 变量                         | $Ret[2,5]$ |            | $Ret[2,5]$ |            | $Ret[2,5]$ |            | $Ret[2,5]$ |            |
|----------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
|                            | 全样本        |            | 高机构投资者组    |            | 低机构投资者组    |            | 低-高机构投资者组  |            |
|                            | 系数         | t 值        | 系数         | t 值        | 系数         | t 值        | 系数         | t 值        |
| $\alpha$                   | 0.038***   | 70.65      | 0.037***   | 71.48      | 0.041***   | 72.53      | 0.005***   | 11.42      |
| $AbRet$                    | -0.112***  | -4.14      | -0.044***  | -3.92      | -0.121***  | -3.87      | -0.077***  | -3.01      |
| $AbRet \times Similarity$  | -0.024***  | -18.41     | -0.023***  | -16.74     | -0.027***  | -19.35     | -0.004***  | -16.82     |
| $Similarity$               | -0.003***  | -4.18      | -0.003***  | -4.15      | -0.005***  | -4.66      | -0.002***  | -3.08      |
| $AbRet \times Ment\_Times$ | -0.023***  | -13.15     | -0.019***  | -10.71     | -0.024***  | -17.32     | -0.005***  | -8.86      |
| $Ment\_Times$              | 0.003***   | 12.88      | 0.002***   | 9.35       | 0.004***   | 18.78      | 0.002***   | 8.73       |
| $AbRet \times Sentiment$   | 0.042***   | 10.25      | 0.041***   | 9.64       | 0.058***   | 16.79      | 0.017***   | 7.74       |
| $Sentiment$                | 0.001***   | 4.81       | 0.001***   | 4.88       | 0.003***   | 5.32       | 0.002***   | 3.76       |
| $AbRet \times MktCap$      | 0.013***   | 9.83       | 0.012***   | 8.91       | 0.019***   | 10.57      | 0.007***   | 5.16       |
| $MktCap$                   | -0.001***  | -58.59     | -0.001***  | -52.18     | -0.003***  | -76.47     | -0.002***  | -12.53     |
| $AbRet[-5,-1]$             | 0.041***   | 62.15      | 0.037***   | 57.82      | 0.048***   | 74.85      | 0.011***   | 16.92      |
| $AbTurn$                   | 0.001***   | 46.73      | 0.001***   | 51.81      | 0.002***   | 56.33      | 0.001***   | 7.49       |
| $Volat$                    | 0.093***   | 68.46      | 0.086***   | 60.63      | 0.135***   | 82.77      | 0.049***   | 36.16      |
| $Illiq$                    | 0.001***   | 15.38      | 0.001***   | 16.14      | 0.002***   | 20.42      | 0.001***   | 9.73       |
| $AbRet \times AbTurn$      | -0.013***  | -28.82     | -0.014***  | -28.84     | -0.018***  | -35.32     | -0.004***  | -4.07      |
| $AbRet \times Volat$       | -0.074***  | -53.63     | -0.062***  | -48.42     | -0.093***  | -58.61     | -0.031***  | -27.04     |
| $AbRet \times Illiq$       | 0.012***   | 8.65       | 0.011***   | 8.35       | 0.016***   | 12.38      | 0.005***   | 3.46       |
| $Ind$                      | Yes        |            | Yes        |            | Yes        |            | Yes        |            |
| F 值                        | 7238.42    | Pr < 0.001 | 8356.57    | Pr < 0.001 | 8671.83    | Pr < 0.001 | 6531.81    | Pr < 0.001 |
| 调整后的 $R^2$                 | 0.326      |            | 0.242      |            | 0.347      |            | 0.226      |            |
| $N$                        | 1046620    |            | 626246     |            | 420374     |            | 420374     |            |

第一,在全样本、高机构投资者组、低机构投资者组中,公司  $i$  在  $t$  日的超额收益率( $AbRet$ )的系数均显著为负,分别为  $-0.112$ 、 $-0.044$  和  $-0.121$ ,这说明网络传播误导性旧信息发布后一周( $[t+2, t+5]$ )收益反转,与大多数关于受市场操纵后公司收益表现的文献观点吻合。此外,在低-高机构投资者组中,超额收益率( $AbRet$ )的系数显著为负,说明低机构投资者组受网络传播误导性旧信息的影响更显著,这与理论模型的推导结果相吻合。

第二,在全样本、高机构投资者组、低机构投资者组中,公司  $i$  在  $t$  日的超额收益率与财经新闻文本相似性交互项( $AbRet \times Similarity$ )的系数均显著为负,分别为  $-0.024$ 、 $-0.023$  和  $-0.027$ ,这表明财经新闻文本相似性越高,则未来一周( $[t+2, t+5]$ )收益反转越明显。该结果与余峰燕等(2012)的观点不同,他们认为好消息重复程度越高,涉及股票的收益越低,并没有涉及收益反转问题的讨论。本文进一步认为不论消息性质好与坏,只要文本相似性高,未来一周收益就会下降,出现反转。同样,在低-高机构投资者组中,超额收益率与财经新闻文本相似性交互项( $AbRet \times Similarity$ )的系数显著为负,也说明低机构投资者组受网络传播误导性旧信息的影响更显著,这与理论模型的推导结果一致。

由此,本文证实了网络传播误导性旧信息发布一周后收益反转,且财经新闻文本相似性越高,则未来一周收益反转越明显的观点,也就是说,网络传播误导性旧信息操纵了公司股价,从而支持了假设 1。

## 2. 网络传播误导性旧信息操纵公司股票交易活动分析

表 5 和表 6 报告了对假设 2 的检验结果。其中,表 5 是关于网络传播误导性旧信息操纵公司股票收益波动性的分析,而表 6 是关于网络传播误导性旧信息操纵公司股票超额交易量的分析。考虑到机构投资者和个人投资者对信息反应的差异,同样将样本整体按机构投资者持股比例分为高机构投资者组和低机构投资者组。

表 5 网络传播误导性旧信息与股票收益波动性

| 变量                    | $ AbRet $  |            | $ AbRet $  |            | $ AbRet $  |            | $ AbRet $  |            |
|-----------------------|------------|------------|------------|------------|------------|------------|------------|------------|
|                       | 全样本        |            | 高机构投资者组    |            | 低机构投资者组    |            | 低-高机构投资者组  |            |
|                       | 系数         | t 值        | 系数         | t 值        | 系数         | t 值        | 系数         | t 值        |
| $\alpha$              | 0.031 ***  | 97.36      | 0.018 ***  | 74.63      | 0.019 ***  | 19.38      | 0.001 ***  | 14.72      |
| <i>Similarity</i>     | -0.003 *** | -3.63      | -0.002 **  | -2.25      | -0.005 *** | -6.49      | -0.003 *** | -2.47      |
| <i>MktCap</i>         | -0.004 *** | -44.56     | -0.003 *** | -42.83     | -0.007 *** | -64.41     | -0.004 *** | -22.87     |
| <i>AbRet</i> [-5, -1] | 0.031 ***  | 97.63      | 0.023 ***  | 96.74      | 0.045 ***  | 106.82     | 0.022 ***  | 35.62      |
| <i>Volat</i>          | 0.384 ***  | 186.44     | 0.386 ***  | 177.25     | 0.571 ***  | 206.16     | 0.185 ***  | 51.35      |
| <i>Illiq</i>          | -0.001 *** | -14.76     | -0.001 **  | -2.12      | -0.002 *** | -23.73     | -0.001 **  | -1.97      |
| <i>Ment_Times</i>     | -0.002 *** | -3.27      | -0.001 **  | -2.14      | -0.003 *** | -8.73      | -0.002 **  | -1.91      |
| <i>Sentiment</i>      | 0.002 ***  | 7.54       | 0.001 **   | 2.37       | 0.004 ***  | 9.83       | 0.003 ***  | 3.36       |
| <i>Ind</i>            | Yes        |            | Yes        |            | Yes        |            | Yes        |            |
| F 值                   | 10542.23   | Pr < 0.001 | 7654.34    | Pr < 0.001 | 7842.77    | Pr < 0.001 | 4252.61    | Pr < 0.001 |
| 调整后的 $R^2$            | 0.206      | 0.197      | 0.335      | 0.174      |            |            |            |            |
| <i>N</i>              | 1046620    | 626246     | 420374     | 420374     |            |            |            |            |

表 6 网络传播误导性旧信息与股票超额交易量

| 变量                    | <i>AbTurn</i> |            | <i>AbTurn</i> |            | <i>AbTurn</i> |            | <i>AbTurn</i> |            |
|-----------------------|---------------|------------|---------------|------------|---------------|------------|---------------|------------|
|                       | 全样本           |            | 高机构投资者组       |            | 低机构投资者组       |            | 低 - 高机构投资者组   |            |
|                       | 系数            | t 值        | 系数            | t 值        | 系数            | t 值        | 系数            | t 值        |
| $\alpha$              | 2.468 ***     | 65.75      | 1.494 ***     | 36.83      | 5.832 ***     | 82.63      | 4.332 ***     | 33.14      |
| <i>Similarity</i>     | -0.122 ***    | -6.62      | -0.088 **     | -2.25      | -0.184 ***    | -10.46     | -0.096 **     | -2.14      |
| <i>MktCap</i>         | -0.061 ***    | -36.46     | -0.038 ***    | -22.85     | -0.107 ***    | -72.17     | -0.069 ***    | -18.56     |
| <i>AbRet</i> [-5, -1] | 0.085 ***     | 30.57      | 0.076 ***     | 29.52      | 0.167 ***     | 44.39      | 0.091 ***     | 13.73      |
| <i>Volat</i>          | -0.084 ***    | -33.73     | -0.065 ***    | -21.93     | -0.136 ***    | -42.17     | -0.071 ***    | -16.59     |
| <i>Illiq</i>          | -0.021 ***    | -17.93     | -0.013 ***    | -11.52     | -0.089 ***    | -26.64     | -0.076 ***    | -8.55      |
| <i>Ment_Times</i>     | 0.144 ***     | 11.14      | 0.111 ***     | 8.46       | 0.223 ***     | 22.65      | 0.112 ***     | 5.53       |
| <i>Sentiment</i>      | -0.021 ***    | -6.36      | -0.016 **     | -2.25      | -0.037 ***    | -13.48     | -0.021 ***    | 4.16       |
| <i>Ind</i>            | Yes           |            | Yes           |            | Yes           |            | Yes           |            |
| F 值                   | 8526.71       | Pr < 0.001 | 3562.66       | Pr < 0.001 | 3284.52       | Pr < 0.001 | 1516.31       | Pr < 0.001 |
| 调整后的 $R^2$            | 0.271         |            | 0.231         |            | 0.284         |            | 0.212         |            |
| <i>N</i>              | 1046620       |            | 626246        |            | 420374        |            | 420374        |            |

由表 5 可以看到,财经新闻文本相似性(*Similarity*)对股票收益波动性( $|AbRet|$ )的影响整体为负,系数为 -0.003,在 1% 的水平下显著,说明财经新闻文本相似性会降低股票收益波动性。分样本来看,在低 - 高机构投资者组中,财经新闻文本相似性(*Similarity*)的系数为 -0.003,在 1% 的水平下显著,说明低机构投资者组中,财经新闻文本相似性对降低股票收益波动性更为显著。

由表 6 可以看到,财经新闻文本相似性(*Similarity*)对股票超额交易量(*AbTurn*)的影响整体为负,系数为 -0.122,在 1% 的水平下显著,说明财经新闻文本相似性会降低股票超额交易量。分样本来看,在低 - 高机构投资者组中,财经新闻文本相似性(*Similarity*)的系数为 -0.096,在 5% 的水平下显著,说明低机构投资者组中,财经新闻文本相似性对降低公司股票超额交易量更为显著。

由此,可以得出结论,网络传播误导性旧信息会降低公司股票收益波动性和股票超额交易量,从而证实了假设 2,且与理论模型的推导结果一致,低机构投资者组强化了网络传播误导性旧信息的影响。

#### (四)稳健性检验

本文还从以下四个方面做了稳健性检验,以进一步检验研究结果的稳健性。

(1)用 Doc2Vec 方法度量财经新闻文本相似性,观察网络传播误导性旧信息操纵市场的情况。为了确保财经新闻文本相似性识别指标选取的稳健性,采用 Doc2Vec 方法度量财经新闻文本相似性。不同于 TF-IDF 法主要识别相同词语的特点,Doc2Vec 方法主要用于识别语义相似的两句话。(2)根据关键词区分非理性和理性,观察网络传播误导性旧信息操纵市场的情况。新闻报道中的信息包含两个方面:一方面是基本面信息,另一方面是噪声信息。以金融词典中特定的表达经济前景的关键词,提取理性的经营状况的传播,进而将总样本按误导性信息区分为理性信息(理性的

业绩相关信息)组和情绪信息(非理性信息)组,再加入所有控制变量进行回归。(3)用不同的时间窗口来观察网络传播误导性旧信息操纵市场的情况。将样本考察区间进一步分为 $[t+1, t+2]$ 、 $[t+1, t+5]$ 以及 $[t+1, t+10]$ 等不同时间段进行回归。(4)应用样本外 2016 年的财经新闻文本相似性数据来考察网络传播误导性旧信息操纵市场的情况。

经过上述稳健性检验,发现结果均与主回归结果一致,支持了假设 1 和假设 2。<sup>①</sup>

## 六、网络传播误导性旧信息操纵市场的机制分析

对于网络传播误导性旧信息操纵市场的机制,可能的解释是投资者行为偏见(Barber 和 Odean, 2008; Tetlock, 2011),即个人投资者容易混淆财经新闻中包含的新信息与旧信息,并且按照旧信息进行交易。为了验证这种可能的投资者行为偏见观点,本文按照公司  $i$  在  $t$  日的超额收益率与财经新闻文本相似性交互项( $AbRet \times Similarity$ )构造对冲组合,即买入低机构投资者组中高的  $AbRet \times Similarity$  组合,卖出低的  $AbRet \times Similarity$  组合,并按照 Fama-French 三因子<sup>②</sup>模型进行回归,以对冲组合的超额收益率作为验证指标。在实证研究中,将样本数据按照机构投资者持股比例分为高机构投资者组和低机构投资者组两组,在低机构投资者组中,按照公司  $i$  在  $t$  日的超额收益率与财经新闻文本相似性交互项( $AbRet \times Similarity$ )分为高组合、中组合、低组合三组,并以高组合组的交互项减去低组合组的交互项,得到  $AbRet \times Similarity$  高减低组合组。在此基础上,运用 Fama-French 三因子模型进行回归。为了保证模型的稳健性,参考 Carhart(1997)的方法加入动量因素( $UMD$ )进行回归,且持有期间除 $[t+2, t+5]$ 以外,还加入了 $[t+1, t+10]$ 期间的收益,结果如表 7 所示。

表 7 基于财经新闻文本相似性的对冲组合(Fama-French 三因子和四因子模型回归)

| 变量         | $AbRet \times Similarity$ 高减低组合组 |            |              |            |               |            |               |            |
|------------|----------------------------------|------------|--------------|------------|---------------|------------|---------------|------------|
|            | $[t+2, t+5]$                     |            | $[t+2, t+5]$ |            | $[t+1, t+10]$ |            | $[t+1, t+10]$ |            |
|            | 系数                               | t 值        | 系数           | t 值        | 系数            | t 值        | 系数            | t 值        |
| $\alpha$   | -0.018 ***                       | -95.23     | -0.016 ***   | -88.62     | -0.016 ***    | -78.22     | -0.015 ***    | -74.52     |
| $Rmrf$     | 0.592 ***                        | 17.82      | 0.484 ***    | 15.28      | 0.426 ***     | 11.05      | 0.28 ***      | 8.03       |
| $SMB$      | -1.122 ***                       | -31.61     | -1.121 ***   | -30.35     | -1.171 ***    | -26.16     | -1.168 ***    | -25.93     |
| $HML$      | 0.168 ***                        | 4.16       | 0.325 ***    | 8.15       | -0.476 ***    | -8.35      | -0.305 ***    | -5.94      |
| $UMD$      |                                  |            | -0.321 ***   | -48.53     |               |            | -0.351 ***    | -43.28     |
| Ind        | Yes                              |            | Yes          |            | Yes           |            | Yes           |            |
| F 值        | 632.62                           | Pr < 0.001 | 983.81       | Pr < 0.001 | 238.257       | Pr < 0.001 | 648.92        | Pr < 0.001 |
| 调整后的 $R^2$ | 0.214                            |            | 0.215        |            | 0.204         |            | 0.212         |            |
| $N$        | 123616                           |            | 123616       |            | 123616        |            | 123616        |            |

① 受篇幅所限,此处仅汇报结论,结果备索。

② 三因子分别为市场资产组合超额收益率( $Rmrf$ )、市值因子( $SMB$ )和账面市值比因子( $HML$ )。三因子数据来自 RESSET 数据库。

在持有期间 $[t+2, t+5]$ 内,对冲组合均获得了显著的负收益。以 Fama-French 三因子模型为例,对冲组合的超额收益率为 $-1.8\%$ ,  $t$  值为 $-95.23$ ,在 $1\%$ 的水平下显著。在加入动量因素(UMD)后,对冲组合的超额收益率为 $-1.6\%$ ,  $t$  值为 $-88.62$ ,同样在 $1\%$ 的水平下显著。当持有期间扩展为 $[t+1, t+10]$ 后,结果没有变化。

本文还对高机构投资者组按上述过程构造对冲组合,结果显示对冲组合在持有期间未能获得显著的负收益。

因此,上述结果有效地支持了个人投资者存在的行为偏见,即个人投资者混淆了财经新闻中包含的新信息与旧信息,并且按照旧信息进行交易。也正是这种认知偏差,使得网络传播误导性旧信息操纵市场现象得以发生。

## 七、结论与启示

市场操纵一直存在,且是世界各国监管机构都非常重视的问题。随着智能互联网时代的到来,以机器学习或神经网络为基础生产和推荐的算法新闻,因算法黑箱过程,难免输出与客观事实不符的结果,误导投资者决策,操纵资本市场,成为金融风险爆发的隐患。当前,要维护金融安全,守住不发生系统性风险底线,有必要加强此类信息型市场操纵研究。

本文运用大数据技术采集并分析主要新闻媒体网站上发布的与 A 股全部上市公司相关的新闻,结合机器学习和自然语言处理技术对收集到的上千万条原始新闻进行严谨的清理分析及核验,在解决新闻与目标公司匹配问题的基础上,构建了财经新闻文本相似性指标。本文以财经新闻文本相似性为网络传播误导性旧信息的代理变量,通过回归分析,证实了网络传播误导性旧信息会操纵资本市场,具体表现在以下几个方面。第一,网络传播误导性旧信息会操纵公司股票收益,在未来一周,公司股票超额收益反转,且财经新闻文本相似性越高,收益反转越明显。第二,网络传播误导性旧信息会操纵对应公司的股票交易活动,即降低公司股票收益波动性与超额交易量,该结论在经过稳健性检验后依然不变。第三,机构投资者持股比例较低的公司受网络传播误导性旧信息操纵的影响更大。第四,网络传播误导性旧信息操纵资本市场的机制在于个人投资者行为偏见,即个人投资者混淆了财经新闻中包含的新信息与旧信息,并且按照旧信息进行交易。

本文的政策启示包含以下几个方面。(1)当前网络科技与资本市场信息服务融合创新发展,必须处理好资本市场发展、稳定和安全的关系到。要落实党的十九届五中全会精神,将网络科技涉及的资本市场信息服务纳入监管,有效防范风险。(2)探索实施监管科技,为建设监管大数据平台提供科学指引。实时对互联网平台算法产生的财经新闻进行监控,设定财经新闻文本相似性安全范围,超过规定阈值的,即可启动市场信息操纵稽查。(3)加强投资者教育。要让投资者意识到以机器学习或神经网络为基础生产和推荐的算法新闻,由于算法黑箱过程而具有极大的欺骗性和危害性。在运用算法新闻进行交易时,要核验信息的客观性与即时性。

### 参考文献:

1. 丁慧、吕长江、陈运佳:《投资者信息能力:意见分歧与股价崩盘风险——来自社交媒体“上证 e 互动”的证据》,《管理世界》2018 年第 9 期。
2. 冯旭南:《中国投资者具有信息获取能力吗?——来自“业绩预告”效应的证据》,《经济学(季刊)》2014 年第 3 期。
3. 孔东民、刘莎莎、谭伟强:《分析师评级与投资者交易行为》,《管理世界》2019 年第 1 期。
4. 李志辉、王近、李梦雨:《中国股票市场操纵对市场流动性的影响研究——基于收盘价操纵行为的识别与监测》,《金融研

究》2018 年第 2 期。

5. [美] 罗伯特·J. 希勒:《非理性繁荣》,李心丹等译,中国人民大学出版社 2007 年版。
6. 游家兴、吴静:《沉默的螺旋:媒体情绪与资产误定价》,《经济研究》2012 年第 7 期。
7. 余峰燕、郝项超、梁琪:《媒体重复信息行为影响了资产价格么?》,《金融研究》2012 年第 10 期。
8. Bagnoli, M. , & Lipman, B. , Stock Price Manipulation through Takeover Bids. *Rand Journal of Economics*, Vol. 27, No. 1, 1996, pp. 124 – 147.
9. Barber, B. M. , & Odean, T. , All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, Vol. 21, No. 2, 2008, pp. 785 – 818.
10. Barberis, N. , Shleifer, A. , & Vishny, R. , A Model of Investor Sentiment. *Journal of Financial Economics*, Vol. 49, No. 3, 1998, pp. 307 – 343.
11. Ben-David, I. , Franzoni, F. , Landier, A. , & Moussawi, R. , Do Hedge Funds Manipulate Stock Price? . *Journal of Finance*, Vol. 68, No. 6, 2013, pp. 2383 – 2434.
12. Bernhardt, D. , & Miao, J. , Informed Trading When Information Becomes Stale. *Journal of Finance*, Vol. 59, No. 1, 2005, pp. 339 – 390.
13. Birz, G. , Stale Economic News, Media and the Stock Market. *Journal of Economic Psychology*, Vol. 75, No. 3, 2017, pp. 87 – 102.
14. Brunnermeier, M. K. , Information Leakage and Market Efficiency. *Review of Financial Studies*, Vol. 18, No. 2, 2005, pp. 417 – 457.
15. Carhart, M. M. , On Persistence in Mutual Fund Performance. *Journal of Finance*, Vol. 52, No. 1, 1997, pp. 57 – 82.
16. Comerton-Forde, C. , & Putniņš, T. J. , Measuring Closing Price Manipulation. *Journal of Financial Intermediation*, Vol. 20, No. 2, 2011, pp. 135 – 158.
17. Daniel, K. , Hirshleifer, D. , & Subrahmanyam, A. , Investor Psychology and Security Market Under-and Overreactions. *Journal of Finance*, Vol. 53, No. 6, 1998, pp. 1839 – 1886.
18. Fama, E. F. , & French, K. R. , The Cross-section of Expected Stock Returns. *Journal of Finance*, Vol. 47, No. 2, 1992, pp. 427 – 465.
19. Fama, E. F. , & MacBeth, J. D. , Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, Vol. 81, No. 3, 1973, pp. 607 – 636.
20. Fruth, A. , Schoneborn, T. , & Urusov, M. , Optimal Trade Execution and Price Manipulation in Order Books with Time-Varying Liquidity. *Mathematical Finance*, Vol. 24, No. 4, 2013, pp. 651 – 695.
21. Gropp, R. , & Kadareja, A. , Stale Information, Shocks, and Volatility. *Journal of Money, Credit and Banking*, Vol. 44, No. 6, 2012, pp. 1117 – 1149.
22. Hong, H. , & Stein, J. C. , A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *Journal of Finance*, Vol. 54, No. 6, 1999, pp. 2143 – 2184.
23. Huang, X. , Teoh, S. H. , & Zhang, Y. , Tone Management. *Accounting Review*, Vol. 89, No. 3, 2014, pp. 1083 – 1113.
24. Khwaja, A. , & Mian, A. , Unchecked Intermediaries: Price Manipulation in an Emerging Stock Market. *Journal of Financial Economics*, Vol. 78, No. 1, 2005, pp. 203 – 241.
25. Kim, K. S. , Park, J. , Chung, C. Y. , & Lee, J. H. , Is Stock Split a Manipulation Tool? Evidence from the Korean Stock Market. *Asia-Pacific Journal of Financial Studies*, Vol. 41, No. 5, 2012, pp. 637 – 663.
26. Lash, S. , Power after Hegemony: Cultural Studies in Mutation? . *Theory, Culture & Society*, Vol. 24, No. 3, 2007, pp. 55 – 78.
27. Li, A. , Wu, J. , & Liu, Z. , Market Manipulation Detection Based on Classification Methods. *Procedia Computer Science*, Vol. 122, 2017, pp. 788 – 795.
28. Loughran, T. , & McDonald, B. , When is a Liability Not a Liability? Textual Analysis, Dictionaries and 10 – Ks. *Journal of Finance*, Vol. 66, No. 1, 2011, pp. 35 – 65.
29. Neupane, S. , Rhee, S. G. , Vithanage, K. , & Veeraraghavan, M. , Trade-based Manipulation: Beyond the Prosecuted Cases. *Journal of Corporate Finance*, Vol. 42, No. 6, 2017, pp. 115 – 130.
30. Peng, L. , & Roell, A. , Managerial Incentives and Stock Price Manipulation. *Journal of Finance*, Vol. 69, No. 2, 2013, pp. 487 – 526.
31. Putniņš, T. J. , Market Manipulation: A Survey. *Journal of Economic Surveys*, Vol. 26, No. 5, 2012, pp. 952 – 967.
32. Schroth, J. , Managerial Compensation and Stock Price Manipulation. *Journal of Accounting Research*, Vol. 56, No. 5, 2018,



pp. 1335 – 1381.

33. Strobl, G. , Earning Manipulation and the Cost of Capital. *Journal of Accounting Research*, Vol. 51, No. 2, 2013, pp. 449 – 473.
34. Tetlock, P. C. , Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance*, Vol. 62, No. 3, 2007, pp. 1139 – 1168.
35. Tetlock, P. C. , Saar-Tsechansky, M. , & Macskassy, S. , More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance*, Vol. 63, No. 5, 2008, pp. 1437 – 1467.
36. Tetlock, P. C. , All the News That's Fit to Reprint: Do Investors React to Stale Information? . *Review of Financial Studies*, Vol. 24, No. 5, 2011, pp. 1481 – 1512.
37. Vila, J. L. , Simple Games of Market Manipulation. *Economics Letters*, Vol. 29, No. 1, 1989, pp. 21 – 26.
38. Wu, M. C. , Huang, Y. T. , & Chen, Y. J. , Earnings Manipulation, Corporate Governance and Executive Stock Option Grants: Evidence from Taiwan. *Asia-Pacific Journal of Financial Studies*, Vol. 41, No. 3, 2012, pp. 241 – 257.

## **The Black Box Algorithm: A Study on Market Manipulation by Spreading Misleading Outdated Information Online**

WANG Chun (Nanjing Audit University, 211815)

**Abstract:** In the era of intelligent Internet, news is produced and recommended based on machine learning, and the general public does not understand its inner workings, hence the term “black box algorithm.” The black box algorithm, though seemingly technology-neutral, produces information that is inconsistent with facts, misleading for investment-decision making, destroys the market pricing function, and easily leads to systemic risks. In this paper, the text similarity of financial news is used as the proxy variable of misleading information spread on the Internet, and the vector space model and the TF-IDF (term frequency and inverse document frequency) method are used to measure the misleading information. Based on 1,046,620 sample entries of news data in the quantified public opinion platform of newspapers and periodicals developed by Datago Technology Co. , Ltd. , it is found that misleading information spread on the Internet will manipulate the capital market. The results are as follows. One week after the release of misleading information, the excess return of stocks reversed, and the greater the text similarity, the more obvious the return reversal was. The misleading information also reduced the stock trading activity, i. e. the volatility of the stock earnings and excess trading volume. After grouping the sample companies by the proportion of institutional investors, it is found that the companies with a lower proportion of institutional investors are more vulnerable to the manipulation of misleading information. The misleading information manipulates the market by taking advantage of the behavior bias of individual investors who could not tell outdated information from new information in the financial news and traded based on the outdated information. This study provides a policy basis for the regulation of algorithm-generated financial news produced by Internet platforms.

**Keywords:** Information Online, Textual Analysis, Financial News, Market Manipulation

**JEL:** G12, G14, G18

责任编辑:非 同