4.金融市场因素

4.1股票价格预测

股市预测作为一直以来金融领域研究的热点，搜索、社交媒体、专题论坛等大数据都在股价预测上有所应用，同时也有许多经济学、统计学以及人工智能的相关算法被应用于其中。下面从股价预测的不同数据来源进行相关研究的梳理。

4.1.1 搜索大数据在股票价格预测中的应用

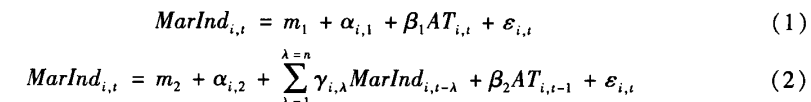
在数据来源上，搜索大数据在股票价格预测中得到了一定的使用。近年来，学术界对金融的投资者关注行为领域产生了浓厚的兴趣，因为股票价格的巨大变化似乎由投资者关注单独驱动。现有研究把投资者关注分为关注( Attention) 和 有限关注( Limited Attention) 。有限关注是指因为投资者的时间和精力有限，不可能考虑所有的股票，对影响股票基本面的信息反应不足，从而对股票市场的价格和成交量产生影响的情况关注有限。关注是指投资者因特定的引人关注的事件，对相应股票产生偏离其基本面的过度反应。对有限关注的研究已经有了很多理论与实证成果，相较而言，因为很难直接观察到投资者关注，所以尽管已经有理论较好地解释了投资者关注对资产价格的影响机制，但实证检验依然困难。

投资者大多通过网络进行股票交易( 特别是在中国①) ，网络用户通常使用搜索引擎来获得信 息。当用户使用搜索引 擎 来 搜 索 某 支 股 票 的 名 称 时，用户对某支股票的关注就直接地显示出来 ( Da et al． ，forthcoming) ，因此，网络总搜索量是投资者关注直接和明确的衡量指标，搜索引擎中金融相关关键词搜索量可代表投资者搜索股市信息行为密集度。这是利用搜索大数据进行股价预测的一个理论依据。

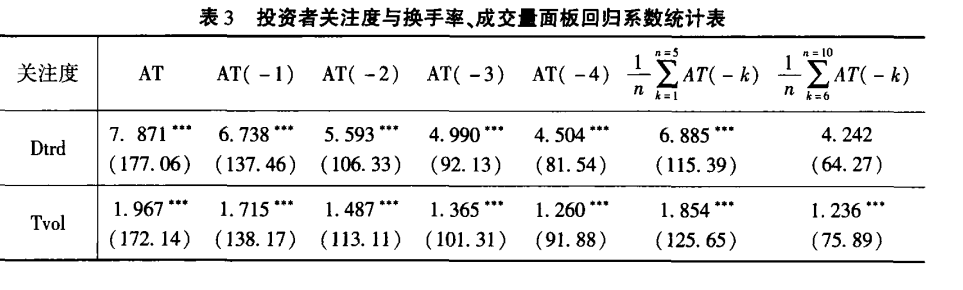
[18]宋双杰, 曹晖, 杨坤. 投资者关注与IPO异象——来自网络搜索量的经验证据[J]. 经济研究, 2011(s1):145-155.用个股公开发行前谷歌搜索量衡量的投资者关注度对于个股市场热销程度、首日超额收益和长期表现有一定的解释力，可以解释首日超额收益的23% ，长期累积收益率的 10% 以上，结果均十分显著。每支股票每 周的搜索量指数。也有研究用与谷歌趋势类似的百度指数构建投资者关注度,发现其对股票市场有正向价格影响[19]俞庆进, 张兵. 投资者有限关注与股票收益——以百度指数作为关注度的一项实证研究[J]. 金融研究, 2012(8):152-165.。其使用的数据为创业板196支股票的百度搜索指数。投资者关注度采用股票简称和股票代码的百度指数相加，取对数得到。



对市场指标和投资者关注度进行了单变量回归，使用模型（1）等验证了百度指数代表的投资者关注度的确与股票市场指标具有显著的相关性，投资者关注度越高，股票的市场表现就越活跃。使用模型（2）在普通的AR模型上增加投资关注指数的滞后项，验证了历史的关注度一定程度上能预测未来的表现，凸显了百度指数区别于市场指标的独特价值。



通过对投资者关注和股票的成交量、换手率的相关性分析，得出了滞后的投资者关注度对上述指标具有显著的正影响。进一步明确了投资者关注本身对股票市场表现有正影响，明确了是投资者先关注某只股票，进行搜索，然后在市场买卖股票，引起了市场指标的变化。而不是某只股票的市场表现突出，引发了投资者的搜索和关注。

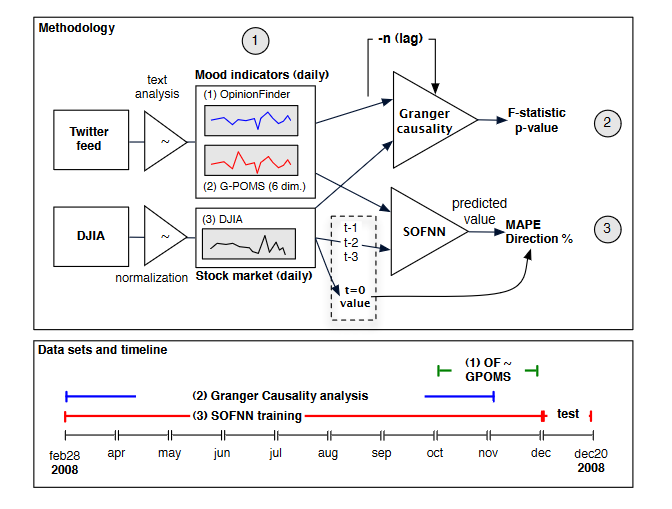


这种投资者关注指标的分析还可以解释更复杂的市场现象，如分析显示，关注度对市场造成正向的价格压力，且这种压力会很快反转，并且非交易日的股票关注与下一交易日集合竞价时的股票价格跳跃显著相关。

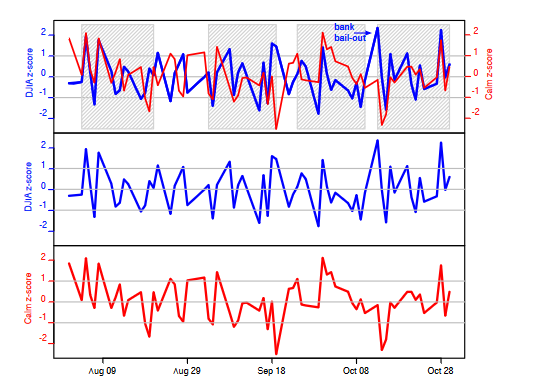
[14] 研究发现金融相关关键词的谷歌搜索量在道琼斯指数涨之前下跌的规律,据此规律建立炒股策略比传统炒股策略收益高。

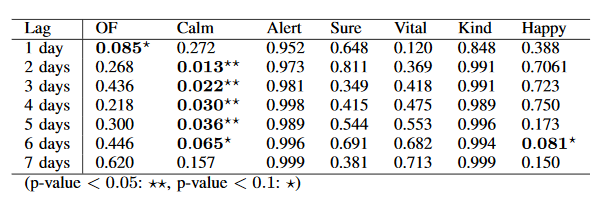
4.1.2 社交媒体大数据在股票价格预测中的应用

微博等社交媒体舆情对预测股价也有显著作用。近些年,以推特和微博为代表的网络社交平台上产生用户生成数据具有空前的规模性和群体性,其中的舆情与股票的市场交易活动也紧密关联。 例如 Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2010, 2(1):1-8.发现基于 Twitter(推特)平台表达的公共情绪可以用来预测股市变动。这一研究的数据来源于两方面：一是 Yahoo·金融发布的道琼斯工业平均指数(DJIA)的收盘价；二是 2008 年 3 月到 10 月，270 万推特用户推送的 970 万条消息。这 970 万条消息经过情绪评估工具——Opinion Finder 和 GPOMS 被赋值。Opinion Finder 根据文本内容可以评估“积极”与“消极”两种情绪；GPOMS 根据文本内容可以评估 “calm(冷静)”、“alert(警觉)”、“sure(确信)”、“vital(活泼)”、“kind(美好)”、“happy(高 兴)”六种情绪，计算出每一天包含某种情绪的微博数量占全部微博的比重，就可以得到衡量公众情绪的时间序列，与股价的时间序列相对应。

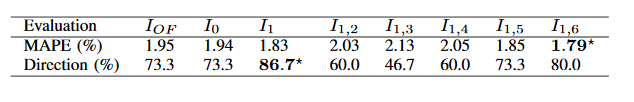


股价与情绪间相关性首先采用加入滞后影响的格兰杰因果检验进行分析，结果发现，在道琼斯工业平均指数（DJIA）和 GPOMS 中的“calm （冷静）”情绪之间存在较为显著的相关性，且情绪对于股价具有一定的预测能力。





进一步股价的预测采用融合了情绪信息的模糊神经网络 Fuzzy Neural Network (SOFNN)。实验发现， “calm（冷静）”情绪可以很好地预测道琼斯工业平均指数在未来 2 到 6 天的涨跌情况，而且这种每日预测的准确率高达到 87.6%。



1. 发现通过挖掘微博文本信息构建的舆情指数可有效预测美国股票价格指数变化。 研究还发现包含观点和情感的微博舆情模型比非舆情微博模型能更准确的预测股票价格变化,熊市预测比牛 市预测更精确[27]。 在中国,“牛市冶”,“利空冶”等股市相关词汇在微博里出现的频率可以短期预测上证指数变化,而负面词汇有更强的预测能力[28] 。

4.1.2 专题论坛与新闻数据在股票价格预测中的应用

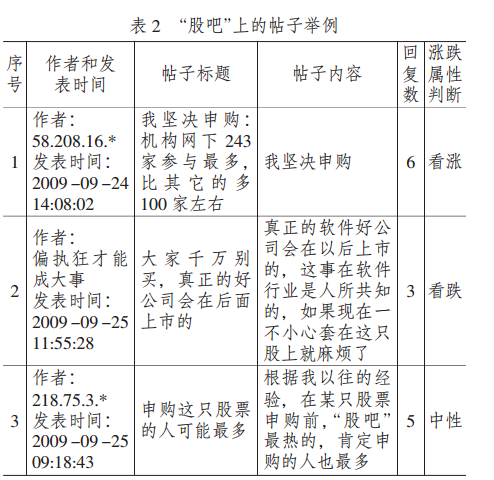
股票专题论坛则是更加集中了投资者感兴趣的话题和讨论，比传统的社交媒体针对性更强。论坛里有关股票收益的讨论比专业分析师的预测更能体现收益预期,根据这种非正式信息进行操盘的收益也更高[22] 。

网络讨论、投资者情绪与 IPO 抑价 林振兴 （厦门大学 管理学院，福建 厦门 361005）

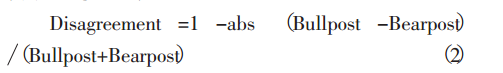
林振兴，2010通过对投资者网络讨论信息进行整理分析，构建了反映投资者情绪的三个指标，即投资者关注度、投资者乐观情绪与投资者意见分歧，研究了投资者情绪与 IPO 抑价之间的关系。

公司首次公开发行的股票通过交易所上市之 后，二级市场的交易价格通常要高于股票的发行价 格，理论界和实务界将这一现象称为 IPO 抑价（IPO Underpricing）。大量的实证研究证实，IPO 抑价现是全球股市普遍存在的现象，其中，马来西亚和中国股市的 IPO 抑价幅度超过了 100%，尤其是中国股市的 IPO 抑价幅度更是高达 256.9%。

采用了86支创业板上市公司的股票价格，以及网络“股吧”的讨论。对目前我国用户访问量最大的互联网财经互动社区之一，我国用户黏性最高的互联网财经互动社区之一——东方财富网旗下频道“股吧”关于创业板个股的发帖讨论进行整理，汇总为投资者情绪指标。

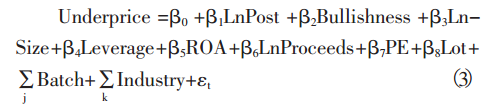


用样本“股吧”在招股 日和公开上市前一日之间的发帖数量来衡量投资者 对于个股的关注程度，人工判断投资者乐观、悲观情绪，并用以下公式来度量投资者的意见分歧（其中，Bullpost 是看涨的帖子数量，Bearpost 是 看跌的帖子数量）。

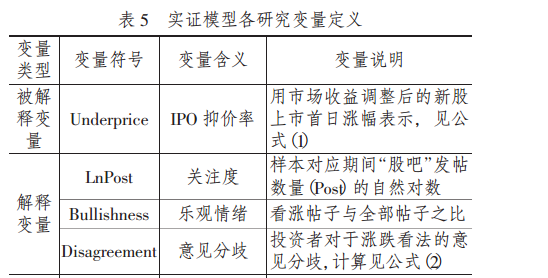


最终建立的实证模型同时控制了样本公司相关财务变量的影响，所有的财务数据均来自 Wind 数据库。包括公司规模（Size，总资产）、财务杠杆（Lever－ age，负债和总资产之比）和盈利能力（ROA，总资产 收益率）；以及新股发行的主要指标，包括募资 金额（Proceeds，实际募资金额）、市盈率（PE，发行价 格与发行后每股净利润之比）以及中签率（Lot，发行 股数和有效申购股数之比）。

建立模型如下，



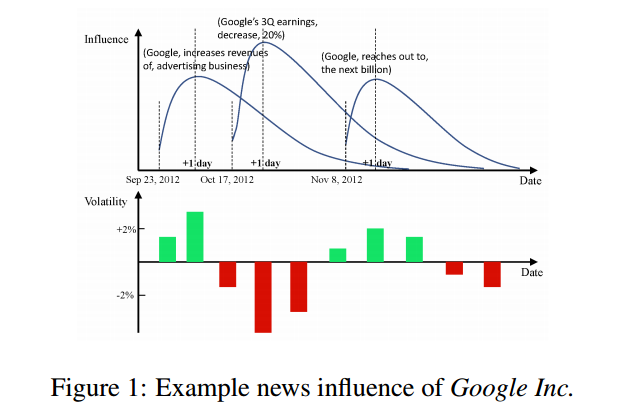
其中解释和被解释变量定义如下，包含了待解释的IPO抑价率，和所研究的情绪指标。

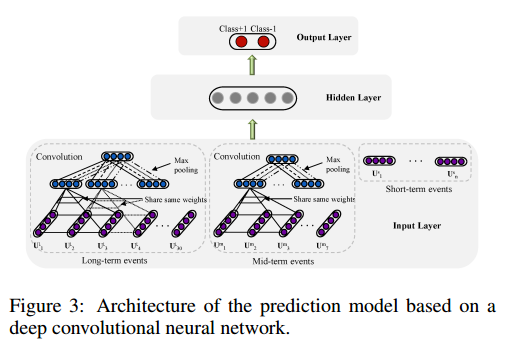


结果表明，应用“股吧”讨论帖子数量作为投资者关注度的代理指标，投资者关注度与 IPO 抑价显著正相关，即投资者关注度越高，IPO 抑价率 越高。并且投资者关注度同样与样本公司 IPO 首日的换手率有显著的正相关关系，即投资者关注度越高， IPO 首日交易量越大。而投资者乐观情绪与投资者意见分歧与IPO抑价和上市首日交易量没有显著关系。

同时, 股票论坛里异常的发帖活动和观点变化与股票异常成交量和回报率有显著关系[23] 。从股票论坛中提取中小投资者观点,构建的投资者情绪指标与传统指标存在一定关联性[25] 。

在新闻数据的应用上，[13]提取新闻词向量并使用 CNN 对 S&P 500 指数及个股价格进行预测， 另外，他们还使用 NLP 技术处理事件文本信息和知识图谱提供的实体、属性等信息预测股市波动率。在 S&P 500指数和个股预测上取得了比现有最好模型6%的提高，模拟实验也显示据此模型建立炒股策略比现有炒股策略收益高。Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:2327-2333

.



4.2 投资市场情绪监测

2012 年，国泰君安推出了“个人投资者投资景气指数”（简称 3I 指数），该指数系国泰君安研究所对海量个人投资者样本进行持续性跟踪监测，对账本投资收益率、持仓率、资金流动情况等一系列指标进行统计、加权汇总后得到的综合性投资景气指数。旨在通过对中小投资人真实投资交易行为的量化解读，更好地了解投资人对市场的预期以及当前的风险偏好等信息。

http://news.cntv.cn/20120711/114745.shtml

　　在样本选择上，选择资金100万元以下、投资年限5年以上的中小投资者，样本规模高达10万，覆盖全国不同地区，具有相当的代表性。对于沪深股市等个人投资者为主的市场更有针对性。在参数方面，主要根据中小投资者持仓率的高低、是否追加资金、是否盈利这几个指标，来看投资者对市场是乐观还是悲观。

“3I指数”每月发布一次，以100为中间值，100—120属于正常区间，120以上表示趋热，100以下则是趋冷。从实验数据看，从2007年至今，“3I指数”的涨跌波动与上证指数走势拟合度相当高。

图为2012年1月至2014年6月3I指数运行走势，可以看出个人投资者的投资景气度处于“低迷”（80以下）。

http://finance.jrj.com.cn/2014/07/07152017555905.shtml



此外，基于上市公司经营报表统计、券商投行研报看涨看跌指数（可按地域/行 业/经营领域细分）的分析，也可以为投资市场情绪提供重要参考依据。

4.3 大宗商品期货相关因素

商品期货是买卖双方在未来某个约定的日期按约定价格进行交易的标准化协议。如何对海量数据进行分析并对期货价格等因素做出预测,以期获取稳定的收益成为商业与学术界关注的重点研究问题。

李新（2015）构建反映投资者对国际原油市场关注度的衡量指标，探讨投资者关注与国际原油市场上不同类型的投资者交易持仓关系。采用网络挖掘-特征选择-广义动态因子模型(WM-FS-FHLR)提取互联网搜索关键词，进行特征选择筛选变量，并构建反映互联网关注的“一致指数”。引入生物信息领域的因果分析方法Partial Directed Coherence(PDC)分析因果关系对所构建的指数进行综合评价。该指数是一个基于互联网数据所构建的综合指数，可用于更全面地衡量原油市场中部分投资者的关注，更及时且准确的表征原油市场的动态变化。

进一步构建多组VAR模型来研究投资者关注指数对非商业投资者、商业投资者及散户的持仓关系，从而分析投资者关注指数与投机活动的领先性。

岳艳涛, 章雅婷, 张宇,等. 基于大数据分析的橡胶期货交易策略研究[C]// 2015.

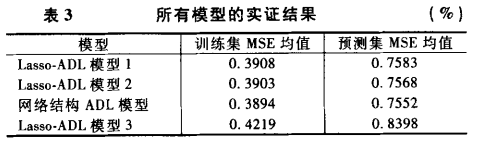
利用期货交易所每秒钟提供两笔交易品种的实时数据(成交量、持仓量、总量、买一价、买一量、卖一价、卖一量)，首先确定影响价格的因素。通过成交价与7个因素的散点图和相关系数分析,发现成交价与B1价S1价相关性较高,然后利用ADF单位根和Johansen检验法对数据原序列、一阶差分序列进行平稳性和协整检验,采用Granger因果关系检验确定具有滞后性的影响因素。

其次对期货价格进行了两部分的预测。首先对橡胶价格进行短期预测,建立自回归与分布滞后模型,用Eviews6.0和SAS9.1软件进行求解和相关性检验。同时利用小波神经网络,对历史数据进行挖掘和模拟分析,进行200和500次训练,预测结果的相关系数达到0.9493,比较预测值与真实值可知模型对短期预测精度高。然后对波动价格长期预测建立阻尼衰减趋势指数平滑模型,并对价格进行预测与误差分析。

进一步，通过期货价格预测还可以建立投资收益模型,以期获得最优收益。从期货交易买卖角度对问题进行分析建立动态规划模型,利用遗传算法,求解预测价格下的交易策略，同样取得了较好的精度和稳定性。

王娜. 基于大数据的碳价预测[J]. 统计研究, 2016, 33(11):56-62.

为了研究大数据是否有助于预测碳排放权价格,本文讨论了结构化数据和非结构化信息对预测碳价所起的作用.结构化数据选取国际碳现货价格、碳期货价格和汇率,非结构化信息选择百度搜索指数和媒体指数.考虑到当解释变量很多时,平等对待每一个解释变量是不合理的,提出了网络结构自回归分布滞后(ADL)模型,在参数估计和变量选择的同时兼顾了解释变量之间的网络关系.实证分析表明,网络结构ADL模型明显优于其他模型,可以获得较高的预测准确性,更适合基于大数据的预测.



从总体上看，Lasso—ADL模型2训练集MSE均值和预测集MSE均值都低于Lasso-ADL模型l，网络结构ADL模型同样如此。LLasso-ADL模型2和网络结构ADL模型都使用了非结构化信息，而LassoADL模型1并没有使用。所以，加入非结构化信息可以改善预测效果。

4.4 其它金融相关因素

对于其它金融相关因素如市场流动性、投资者交易行为等也有利用大数据展开的相关研究，大数据对于金融市场运行规律的揭示起到了一定作用。

王晓琳. 金融投资者行为及其对股票与期货市场的影响研究[D]. 哈尔滨工业大学, 2016.

关于投资者行为如何影响金融市场进行了探讨。首先，探讨了投资者关注对分析师盈余预测行为的影响，运用标准普尔500股票的谷歌搜索趋势作为投资者关注的代理变量，发现搜索量与分析师预测修正、分析师总人数和分析师预测分歧相关。其次，发现我国股指期货市场流动性日内效应与周内效应，得出了期货市场流动性的日内模式在上午呈现反“J”形变化模式，下午呈现倒“U”型变化模式，期货流动性的周内模式呈现倒“V”形变化模式。表明中国股票指数期货市场流动性的周内模式和日内模式存在着显著的周期性。并且，基于指令失衡这一关键指标作为投资者交易行为的代理变量，分析了投资者交易行为与价格行为之间的关系，结果发现在中国农产品期货市场指令失衡具有收益预测的能力。过度指令失衡引起的价格行为并不符合传统库存理论的解释，中国期货市场的价格行为主要由投机者的羊群效应导致。

运用多种统计学与计量经济学中常用的分析方法，包括：时间序列回归分析、协整检验、VECM模型估计、格兰杰因果检验与Fama-MacBeth截面回归模型等，具有前沿性、系统性、市场性和监管导向性。