1. **引言:**

过去许多文献已实证发现投资人的情绪会显著影响股票市场报酬。Hirshleifer and Shumway (2003)实证发现由一个国家城市早晨阳光衡量的愉快投资人情绪会和当日市场指数报酬显著相关。Lemmon and Portniaguina (2006)探讨介于情绪及小股异常报酬的时间序列关系，并发现消费者信心能预测小股报酬。Baker and Wurgler (2006)检测投资人情绪的横断面影响，结果显示当情绪较低(高)时，对小股、年轻股、高波动股及贫困股的后续报酬会相当高(低)。Dash and Maitra (2018)使用回归及小波模型及周数据检测印度股市投资人内含情绪指数及价值加权股市指数之间的关系，结果呈现不论于短期或长期，投资人情绪对股票报酬存在强烈正向的影响。这些文献结果多显示投资人情绪和当期股票报酬有正向关系，而对预期股票报酬有负向关系 (Baker and Wurgler, 2006; Baker et al., 2011; Schmeling, 2009; Dash and Maitra, 2018)。进而,有些文献已利用新闻的文本分析来探讨这个主题。Sun, Najand and Shen (2016)探讨由新闻媒体计算的高频投资人情绪和股票报酬之间的预期关系，结果发现甚至落后报酬存在之下，落后半小时的投资人情绪会最优势的预测日内S&P500指数报酬。Renault (2017)使用一个大众于股市的微博平台，建构由在线大众于分享股市多头或空头看法时使用的字词并形成情绪权重，发现前半小时投资人情绪变动可预测后半小时S&P500指数ETF报酬。然而这些文献的信息多来自单一新闻来源，尚需要多元的信息平台，才能有效结合来自不同投资客群的社会聆听。

过去文献上使用投资人情绪来预测股票报酬率多使用三种方法来衡量投资人情绪。第一种为使用“调查基础”的投资人情绪指数衡量(Brown and Cliff, 2005)，如AAⅡ投资人情绪调查。第二种为“市场基础”的变量。Baker and Wurgler (2006; 2007)使用如IPO第一天报酬、IPO量及交易量以衡量投资人情绪。Brown and Cliff (2004), Smales (2017) and Dash and Maitra (2018) 使用卖权相对买权数目的比率、市场周转率相对股市总市值比率、隐含波动度指数 (VIX) 以衡量投资人情绪。第三种为使用“传统媒体新闻基础”的投资人情绪衡量(Tetlock, 2007; Sun et al., 2016)。Da et al. (2015)提及当调查无反应率高时，调查基础衡量变成不可靠；市场基础衡量可能夹杂经济推动的结果。传统媒体新闻基础通常不是来自单一来源(Tetlock, 2007)，就是混杂股市之外来源的财务情绪数据形成的指数(Sun et al., 2016)；前者获得的情绪衡量不够客观，后者则可能捕捉投资人于股市以外的情绪。

近年来行为财务学者已构建来自网络数据的投资人情绪指针(Renault, 2017; more….)。Renault (2017)提及来自网络的投资人情绪信息应可克服1. 来自于调查基础衡量可能夹杂和答案相关的偏差，2. 来自于市场基础衡量可能有特质及非情绪相关的成分，与3. 来自于传统媒体新闻基础可能混杂因果关系。因为网络沟通的持续变动特质，新闻噪声的问题可被频繁地重新评估。近年一些文献已使用来自网络的信息以预测股票市场报酬，但他们的网络信息来源多来自于如StockTwits的单一网络平台(Da et al., 2015; Chen et al., 2014; Renault, 2017) 有可能搜集到特定客群的情绪反应，需要获得来自不同投资客群的情绪反应才能结合社会聆听。并且目前文献未能直接捕捉投资人对某只股票的情绪，仅能捕捉投资人对整体股市看多或看空的时间序列趋势（如Renault, 2017）。若能掌握投资人对个股乐观及悲观的情绪，可落实择股及择时的投资决策。同时，若能透过网络信息进一步掌握投资人对个股乐观及悲观情绪的强度，更能应用投资人对个股乐观及悲观情绪的网络声量以预测对股票报酬率的影响。

近年来已有许多学者应用金融科技于财务相关的研究。Lu, Shen and Wei (2013)利用计算机语言文本挖掘分析方法对台湾财务新闻进行提取和量化，以期进一步发展经典的财务危机预警模型。并提出"违约主体的危难强度"(DIDC)，通过将该方法纳入信号模型，以及传统的财务绩效变量(流动性、债务、活动性和盈利能力比率)，来研究该方法对违约概率的预测能力。刘海飞，许金涛（2017）结合文本挖掘技术和事件研究方法研究发现: 五种互联网异质性新闻均会引起股票短期的异常收益，其中，政策扶持类、兼收并购类、再融资类和盈利能力类新闻能够对公司股票产生正的异常收益，而违规处罚类新闻能够对公司股票产生负的异常收益。

在实务上，随着相关技术的成熟，证券产业也开始出现应用投资人情绪大数据分析以预测大盘指数股市报酬的案例。例如2010年美国印第安纳大学的Johan Bollen与Huina Mao等人利用分析Twitter发文的情绪，可成功预测87.6%的美国道琼工业指数涨跌趋势。2013年英国华威商学院的Tobias Preis等人发现利用分析Google的关键词搜寻次数，可以预测股市走向。国外市场上亦已有业者推出实时分析网络文章、社群媒体、新闻等讯息以做为投资参考的服务，例如美国The Stock Sonar及我国的情绪宝等。这些实际案例说明结合社群聆听及金融科技方式可捕捉投资人情绪信息，进而能准确预测股市指数的报酬。中国资本市场自2014年起陆续推出大数据指数基金，指数选股标准除了市值、盈余等传统因子外，还加入了在搜索引擎的访问次数及社群媒体的评论正负项得分等大数据因子。

随着互联网的不断发展，当前中国国内网民人数已经超过了人口的一半，与传统的信息传播方式相比，互联网使得网民个体之间的沟通具有一种低成本、传播及时和交流广泛的特点，方便了投资者之间的交流，也使得投资者情绪得以更加广泛的传播（孙鲲鹏，2018）。与传统媒体相比，互联网论坛的信息数量庞大，内容杂乱，如何获取数据并按研究目的对数据进行分类就成为了一个至关重要的问题，而金融科技的引入恰好能解决这一问题。金融科技可以基于网络爬虫技术进行数据的搜索和抓取，并通过文本分析方法对获取的数据进行分类和量化，因此通过网络论坛来捕捉投资人的情绪信息需要利用金融科技。

2017年，中国透過網路交易的数字经济总量达27. 2万亿元，对GDP增长贡献率达55%，全球数字经济规模达12. 9万亿美元，中国位居全球第二。社交平台作为数字经济呈现形式之一，现已是消费者或投资者交换观点、情感和知识的重要渠道。基于庞大的数据量，股吧信息可以比较准确、实时地反映出投资者的心理和行为的变化。与调查问卷、档案数据和访谈记录等信息源相比，社交平台数据能够规避传统信息收集方式的滞后、缺失、主觀和高投入等弊端，具有用户基数大、社交性强、涉入性高、客觀性高與响应速度快等优势。由于数据量庞大，个人投资者人工从冗杂的信息中分辨出所需的真实信息需要付出较髙的时间成本以及大量的脑力劳动，但是金融科技的使用就可以很好地解决这个问题。我们可以使用网络爬虫技术爬取各网络平台的信息，快速、准确地提取到和研究相关的词汇，并在此基础上对获得的数据进行分类汇总和量化。因此我们可以利用金融科技在各网络平台或者新闻媒体中捕捉投资人的情绪并对其进行相关研究。

由于国内股票市场发展的起步较晚，相关机制及监管体系还不够健全，因此现阶段国内股票市场与国外还存在着诸多不同之处，正如当前国内股票市场无法引进熔断机制，而且近年來中国股市的异动完全独立于其他国家股市，波动幅度非常之大。顯示国外的许多研究成果在国内也并不适用。同时，加之互联网社交媒体技术近年在中國大陸快速发展，越来越多的投资者倾向于在股票网络论坛上表达自己对股票或股市的看法和预期，因此利用金融科技研究投资者情绪对股票报酬的影响是很有必要的。

大部分的国外学者研究发现投资者情绪对当期大盘的股票报酬会产生正向影响，而对预期大盘的股票报酬会产生负向影响(Baker and Wurgler, 2006; Baker et al., 2011; Schmeling, 2009; Dash and Maitra, 2018)。然而**国内对投资人情绪和股票报酬的正负向关系却存在着争议。**国内学者段江娇，刘红忠，曾剑平（2017）实证結果和國外結果相似。但是国内学者也在相关研究中发出了其他的声音。王美今,孙建军（2004）发现：投资者情绪变化不仅显著地影响沪深两市的均衡收益, 而且显著地反向修正沪深两市的收益波动。金雪军（2013）认为投资人看涨指数的当期和滞后一期都与股票的收益率正相关。然而俞庆进（2012）证实：滞后4期的投资者关注度对当期收益率有显著地负向影响，段江娇（2014）则认为投资者情绪指数的上涨会导致股指收益率当期和滞后3期之内的不断下降。综上，可看出在借鉴国内外学者研究成果的基础上，利用金融科技获取网络论坛上投资人情绪信息，并结合国内股票市场实情来进行预测股票报酬的研究分析。因此投资人可以通过金融科技大量利用过去在网络论坛上投资人情绪的信息，以精确地预测股市报酬进而获利，对投资人在股市获利有着非常重要的现实意义。

然而，应用金融科技捕捉对股票报酬影响的文章，尚缺乏从投资人对个股情绪的面向深入分析对股票价量的影响。相关文献需进一步利用金融科技直接衡量投资人对个股正面及负面情绪，将正负面情绪置入模型中分析其对股票报酬的影响，才能让投资人在择时之际同时具备择股的能力。因此本研究进而从金融科技视角结合社群聆听，来探讨投资人乐观及悲观情绪是否及如何影响进而预测股票报酬率。

[1] 王美今,孙建军.中国股市收益、收益波动与投资者情绪[J].经济研究, 2004(10): 75-83.

[2] 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(08):152-165.

[3] 段江娇,刘红忠,曾剑平.投资者情绪指数、分析师推荐指数与股指收益率的影响研究——基于我国东方财富网股吧论坛、新浪网分析师个股评级数据[J].上海金融,2014(11):60-64.

**1.2 本课题相对于已有研究的独到学术价值和应用价值**

本课题通过网络爬虫技术爬取各网络社群平台，直接捕捉投资人对某只股票涨跌看法的信息，形成乐观及悲观情绪的网络声量，并将正负面情绪的强度勾稽至金融大数据平台，探讨正负面情绪强度对股价报酬的影响进而预测报酬。

1] 刘海飞,许金涛.互联网异质性财经新闻对股市的影响——来自中国互联网数据与上市公司的证据[J].产业经济研究,2017(01):76-88.

1. **文献综述:**

**2.1 投资人情绪对股票报酬的影响相关文献**

(加入富翔已找的其他文獻)

Fama在有效市场假说（EMH）中，将股票市场建模为随机游走的，从而判断出股票价格是不可预测的。然而，行为金融学在对EMH的批判性解读中表明了股票市场上存着着某种程度的可预测性(DellaVigna, 2009)。过去许多文献已实证表明投资人的情绪会对股票市场报酬产生一定的显著影响。Hirshleifer and Shumway (2003)实证发现由一个国家城市早晨阳光衡量的愉快投资人情绪会和当日市场指数报酬显著相关。Lemmon and Portniaguina (2006)探讨介于情绪及小股异常报酬的时间序列关系，并发现消费者信心能预测小股报酬。Baker and Wurgler (2006)使用了一些简单的理论知识以及关于投机事件的历史记录，研究了投资者情绪对股票报酬的横截面的影响，他们发现当情绪的期初指标较低时，小型股、年轻股票、高波动性股票、无盈利股票、无股利股票、极端成长型股票和不良股票的后续回报率相对较高；当情绪高涨时，这些股票的后续回报则相对较低。随着进一步研究，Baker and Wurgler (2007)提出的投资者情绪方法是“自上而下”和宏观经济学方法，并得到结论：对投资者情绪最敏感的是小股、年轻股、高波动股及贫困股，而且情绪波动对个别公司和整个市场都有明显的、重要的和有规律的影响，尤其对难以套利或估值的股票影响最大。Tetlock(2007)研究媒体内容与股市活动时发现较高的媒体悲观情绪会对市场的价格产生下行压力，同时过高或者过低的悲观情绪也会致使市场交易量暂时偏高。除此之外，悲观情绪对价格的影响在小型股中表现的更为明显，而且逆转速度也较缓慢。Leung and Ton (2015)调查了2003年1月至2008年12月超过250万条HotCopper信息对澳大利亚股市的影响，发现董事会信息和信息情绪的数量与表现不佳(低净资产收益率、EBIT利润率、每股收益)的具有高市场增长潜力(低账面市值)的小盘股的短期回报显著正相关，而大盘股则不会受到留言板活动的影响。胡昌生，迟阳春（2013）研究了在不同估值水平下投资者情绪对于股票市场波动性影响的差异。我们发现在市场不同的估值阶段投资者情绪对波动性的影响有很大差异。当市场 处于高估值期时非理性情绪对波动性有显著影响理性情绪的影响不显著；当市场处于低估值期时，理性情绪对波动性有显著影响，非理性情绪的影响不显著。此外，理性情绪对大盘股波动性的影响更大，非理性情绪则对小盘股波动性的影响大。

近年来有些文献已采用高频率的日数据或日内数据分析投资人情绪与股票市场指数回报率之间的关系。Da Engelberg and Gao (2015)在构建恐惧指数来预测市场回报时提出恐惧指数与当日的低回报率相关，但是可以预测明日的高回报率，这种效应在受投资者青睐且难以套利的股票中表现的尤为显著。此外，恐惧指数与日波动率的短暂成分、VIX期货回报率也密切相关。Dash and Maitra (2018)使用回归及小波模型及周数据检测印度股市投资人内含情绪指数及价值加权股市指数之间的关系，结果呈现不论于短期或长期，投资人情绪对股票报酬存在强烈正向的影响。Gao et al. (2017）利用1993年至2013年的高频标准普尔500交易所交易基金 (ETF) 数据，成功的通过衡量市场开盘时的第一个半小时的市场回报预测了当天市场收盘前最后半小时的回报，并证明这种可预测性在股价波动较大的日子、成交量较高的日子、衰退日和重大宏观经济新闻发布日都比较有效。

许多国外相关文献结果多显示投资人情绪和当期股票报酬有正向关系，而对预期股票报酬有负向关系 (Baker and Wurgler, 2006; Tetlock, 2007; Baker et al., 2011; Da et al., 2015; Schmeling, 2009; Dash and Maitra, 2018)。如Baker and Wurgler (2006)使用美国股市的月数据研究投资者情绪对股市报酬的横截面的影响，发现情绪的期初指标与未来股市指数的回报率呈反向关系。Tetlock(2007)使用道琼斯工业指数的日交易数据研究发现较高的媒体悲观情绪会对当前市场的价格产生下行压力，但在接下来的几天内市场活动又会发生反转。Da Engelberg and Gao (2015)使用日频率的投资者恐惧指数数据构建预测美国市场回报时发现：恐惧指数与当日股票的低回报率相关，但是可以预测明日股票的高回报率。Dash and Maitra (2018)使用周频率的数据分析了印度股市投资人内含情绪指数和价值加权股市指数之间的关系，发现不论短期或长期，投资人情绪对股票报酬均存在强烈的正向影响。然而国内在这方面的研究结果却存在着争议，金雪军（2013）抓取创业板的每日帖子数和内容构建投资人看涨指数和意见趋同指数发现投资人的看涨指数和股票的收益率正相关，看涨指数每增加1个单位，个股收益率就增加个0.7百分点，而且滞后一期的看涨指数对于收益率具有同样的预测作用。段江娇（2014）抓取上证A股的每日帖子内容构建投资者情绪指数，认为投资者情绪指数的上涨会导致股指收益率当期和滞后3期之内的不断下降。由于发展中国家股市和发达国家股市的相关机制不同，市场结构有所差异，股市成熟度不一，所以在相关研究方面的结果也有所不同。因此，投资者情绪和中国股市之间的影响关系是值得我们进行研究的。

DellaVigna, S. (2009) Psychology and economics: evidence from the field. Journal of Economic Literature,47(2): 315–372.

Baker, M. , Wurgler, J. , 2006. Investor sentiment and the cross-section of stock re- turns. J. Finance 61 (4), 1645–1680 .

Gao, L., Han, Y., Li, S. Z., Zhou, G., 2017. Market intraday momentum. Working Paper, Washington University in St. Louis.

Leung, H. , Ton, T. , 2015. The impact of internet stock message boards on cross-sec- tional returns of small-capitalization stocks. J. Bank. Finance 55, 37–55 .

Doms, M., and N. Morin, 2004, Consumer Sentiment, the Economy, and the News Media, working paper, Federal Reserve Bank of San Francisco.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. J. Econ. Perspect. 21, 129–157.

Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62 (3), 1139–1168.

Da, Z., Engelberg, J., Gao, P., 2015. The sum of all FEARS investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32.

Dash, S., Maitra, D .,2018.Does sentiment matter for stock returns? Evidence from Indian stock market using wavelet approach[J]. Finance Research Letters. 26, 32-39.

* 1. **传统的投资人情绪衡量**

过去的文献关于投资人情绪衡量的方法主要有三种。第一、调查基础衡量。Brown and Cliff (2005)以调查的投资人情绪数据为基础，研究投资人情绪调查数据与长期回报和资产价值之间关系，发现了高情绪之后往往伴随着低收益。第二、市场基础衡量。Baker and Wurgler (2007)则以市场数据为基础，通过计算IPO和投资人情绪之间的关系，证明了衡量投资人的情绪是完全可行的；投资人的情绪对个别公司和整个股市有着十分明显、重要和规律的影响，特别是一些难以套利的股票受投资者情绪影响最大。Liao et al. (2011) 采用三项式分布方法来度量管理者羊群行为，并使用主成分分析作为提取市场基础相关变量的复合情绪度量的手段，证明了投资人情绪在共同基金羊群行为中起着重要作用。第三、传统媒体基础衡量。还有一种以传统新闻媒体为基础的投资人情绪衡量方法，Tetlock (2007)利用《华尔街日报》一篇流行专栏的日常内容，来定量衡量媒体与股市互动的性质，发现媒体的高悲观情绪预示着市场价格的下行压力，下行之后市场价格会再回归到基本面，异常高或异常低的悲观情绪预示着更高的市场交易量。Sun et al. (2016) 探讨了高频投资人情绪与股市收益率之间的预测关系，通过对来自新闻联播、互联网新闻和社交媒体等消息来源的文本进行分析计算，证明了标准普尔500指数的回报率是可以预测的，并且认为回报的可预测性很可能是由噪音交易者的交易活动所驱动的。

* 1. **网络或结合金融科技的投资人情绪衡量**

近年来，越来越多的行为财务学者开始以网络信息**或结合金融科技**为基础衡量投资人的情绪。来自网络的投资人情绪信息应可克服来自于调查基础衡量可能夹杂和答案相关的偏差、市场基础衡量可能有特质及非情绪相关的成分，与传统媒体新闻基础可能混杂因果关系 (Renault, 2017)。同时因为网络沟通的持续变动特质，新闻噪声的问题可被频繁地重新评估。进而，结合金融科技技术可使得用庞大的网络信息以更精确的衡量投资人情绪。

Da et al. (2015) 使用来自数百万家庭的每日互联网搜索量来揭示市场层面的情绪，通过汇总与家庭问题 (例如，“衰退”、“失业”和“破产”) 相关的查询量，构建了一个由金融和经济态度 (恐惧)指数，作为衡量投资者情绪的新指标。Kim and Kim (2014) 从雅虎财经留言板上下载了2005年1月至2010年12月期间关于91家公司超过3200万条的信息，从中提取投资者的情绪，根据散户投资者明确的情绪变化来构建投资者情绪指数。Kearney and Liu (2014)指出越来越复杂的文本内容分析加上更广泛的特定领域词典，为我们提供了更准确、更有效的情绪度量。Oliveira et al. (2016）基于统计方法从 Stocktwis平台获取了大量的信息，提出了一种新颖且快速的股票市场词汇创建方法，生成了一种更为有效的投资人情绪衡量指标，并且指出新的网络搜索方法相对传统的调查指标更具有优势。

少数国内学者也摘录网络信息以衡量投资人情绪。例如，王靖一 (2018）则收集了和讯网2013年1月到2017年9月1702多万条新闻数据编制了北京大学金融科技情绪指数，数据规模超过了700GB。在应用金融科技来捕捉投资人情绪方面的研究起步较晚，但近年来也取得了一些成果。唐涛提出了当前网络情绪的分析已经超过了传统的分析框架，必须结合金融大数据在分析方法上进行创新，并且指出未来的网络情绪分析方法应该向金融大数据分析的方向发展创新（唐涛，2014）。李金海(2014)根据大数据思想构造了网络情绪的文本挖掘模块，通过计算词汇在文件中的重要性来构造情绪指数。孟雪井等学者曾使用过文本挖掘技术来对中国知网的CSSCI期刊以及新浪微博上的话题进行文本分析，并结合了百度的关键词推荐系统，综合了三大词库来确定我国与投资者情绪相关的网络关键词（孟雪井，2016）。王夫乐等学者通过新浪微博开发平台的开放接口，抓取了微博上2013年3月25日到2016年2月26日的每日数据，并对每一条微博内容都进行了文本情感分析，将情感分为喜悦、惊奇、愤怒、恐惧、悲伤五个维度，并以每一种情绪的微博数量作为权重，以各种情绪的加权均值来代表当日的总情绪（王夫乐，2017）。杨欣等学者利用从百度指数获得的搜索量数据，构建了衡量投资人对于突发事件的关注程度的指标。

1. Antweiler, W. , Frank, M.Z. , 2004. Is all that talk just noise? the information content of internet stock message boards. J. Finance 59 (3), 1259–1294 .
2. Baker, M. , Wurgler, J. , 2007. Investor sentiment in the stock market. J. Econ. Per- spect. 21 (2), 129–152
3. Brown, G.W. , Cliff, M.T. , 2005. Investor sentiment and asset valuation. J. Bus. 78 (2), 405–440 .
4. Chen, H. , De, P. , Hu, Y.J. , Hwang, B.-H. , 2014. Wisdom of crowds: the value of stock opinions transmitted through social media. Rev. Financ. Stud. 27 (5), 1367–1403 .
5. Da, Z. , Engelberg, J. , Gao, P. , 2015. The sum of all FEARS: investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32 .
6. Das, S.R. , Chen, M.Y. , 2007. Yahoo! for amazon: sentiment extraction from small talk on the web. Manag. Sci. 53 (9), 1375–1388 .
7. Liao,T.,Huang, C.,Wu,C., 2011. Do fund managers herd to counter investor sentiment?[J]. Journal of Business Research ,64(2).
8. Kearney, C. , Liu, S. , 2014. Textual sentiment in finance: a survey of methods and models. Int. Rev. Financ. Anal. 33 (3), 171–185 .
9. Kim, S.-H. , Kim, D. , 2014. Investor sentiment from internet message postings and the predictability of stock returns. J. Econ. Behav. Organ. 107, 708–729 .
10. Oliveira, N., Cortez, P., Areal, N., 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis. Support Syst. 85, 62–73 .
11. Sprenger, T.O. , Sandner, P.G. , Tumasjan, A. , Welpe, I.M. ,2014. News or noise? using twitter to identify and understand company-specific news flow. J. Bus. Finance Account. 41 (7–8), 791–830 .
12. Sun, L., Najand, M., Shen, J., 2016. Stock return predictability and investor senti- ment: a high-frequency perspective. J. Bank. Finance 73, 147–164 .
13. Tetlock, P.C. , 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62 (3), 1139–1168 .
    1. **使用网络社群聆听或结合金融科技捕捉投资人情绪信息並预测股票报酬 (请富翔找英文相关文献)**
14. (1) 相關英文文獻

多数网络社群聆听倾向捕捉大盘股市的投资人情绪信息，进而分析投资人情绪对股市报酬的影响。Moat et al. (2013) 认为投资者都是风险厌恶者，并对2007年12月至2012年4月间维基百科的历史使用数据进行了研究，发现在股市下跌之前，与公司或其他金融话题相关的文章的页面浏览量会增加，从而验证互联网上的投资人情绪信息搜集过程会领先于现实世界的数据变化。使用网络信息衡量投资人的情绪时，由于信息量十分庞大，传统的信息处理方式已经越来越不能满足工作的需要。使用金融科技一方面可以挖掘更多投资人信息以提高投资人情绪衡量的精确度；另一方面，可同时精确估计投资人情绪衡量对股票报酬率的影响。缘此，现在的行为财务学者在数据处理上开始使用金融大数据或金融科技技术相关的创新。Antweiler和Frank (2004) 开拓性的使用了互联网留言板上的数据，他们搜集了雅虎上150多万条信息，经过测试发现这些信息可以预测股票市场的波动。Das and Chen (2007) 由投票选择出一些结合在一起的分类器算法，开发出一种从股票留言板中提取小投资人情绪的算法。精度水平类似于广泛使用的贝叶斯分类器，但误报较低，情绪精度较高。并通过该算法证明了技术部门发布的信息与股票指数水平以及成交量和波动性有关。Ranco等学者（2015）追踪了15个月中，形成道琼斯工业平均指数 (DJIA) 的30家股票公司的推特数量和情绪，并利用监督机器学习的方法对推特上的情绪进行了自动分类。研究发现，在相关推特数量的高峰时期，推特情绪与异常回报之间存在着明显的依赖关系，推特高峰时期的极端情绪暗示了累积异常回报的方向。Bollen，Mao and Zeng（2011）调查从大规模推特反馈中得出的集体情绪状态的测量值是否与道琼斯工业平均指数(DJIA)随时间而变化的价值相关。他们通过两种情绪跟踪工具来分析每日推特的文本内容，测量情绪的谷歌情绪状态档案(GPOMS)。然后检测公众对2008年总统选举和感恩节的反应的能力来交叉验证情绪时间序列的结果。结果表明，通过纳入特定的公众情绪维度，而不是其他维度，DJIA预测的准确性可以得到显著提高，预测DJIA收盘价每日上下变化的准确率为87.6%，平均百分比误差减少了6%以上。Kim and Kim (2014) 从雅虎财经留言板撷取散户投资者明确的情绪变化来构建投资者情绪指数，并采用了Naive Bayes分类算法进行分类，来考察投资人情绪对股票回报的可预测性。Chen et al. (2014) 通过对美国一个十分受欢迎的社交媒体平台上发表的文章进行文本分析，研究了社交媒体传达的投资人意见在多大程度上预测了未来的股票回报。Sprenger et al. (2014) 提出了一种方法来识别基于现实事件的微博信息，并用计算语言学研究了超过40万条投资人对标准普尔500指数相关的推特信息，通过划分好消息和坏消息，证明了好消息发生之前的指数回报率比坏消息发生时更为明显。Sprenger, Tumasjan, Sandner and Welpe（2014）使用類似的作法發現高于平均水平投资用户的建议被转发的频率更高，拥有更多的关注者，从而扩大了他们的声音份额。Renault (2017)使用一个大众于股市的微博平台，建构由在线大众于分享股市多头或空头看法时使用的字词并形成情绪权重，发现前半小时投资人情绪变动可预测后半小时S&P500指数ETF报酬。

国内对于投资人情绪方面的研究获得了许多成果，但研究结果主要集中于情绪影响模型的构造上，然而近年来已有一些文献应用金融科技来捕捉投资人情绪对股票报酬的影响。杨欣 (2014)构建投资人对于突发事件的关注程度的衡量，并以该衡量来研究突发事件关注度对于股数波动的影响。杨晓兰、沈翰彬和祝宇（2016）以投资者在东方财富网股吧针对创业板上市公司发表的 90 多万条帖子为研究对象，通过IP地址识别，构建本地关注指标；并利用计算机文本挖掘技术，提取网络发帖所体现的情绪倾向，构建投资者情绪指标。实证结果显示，本地关注对股票收益率的影响取决于投资者情绪，当投资者持积极情绪时，本地关注对股票收益率有显著的正向影响; 当投资者持消极情绪时，该影响显著为负。在积极情绪和消极情绪下本地关注对股票交易量都有显著正向影响，但积极情绪下的影响程度比消极情绪下更大。石勇等人（2017）基于优矿金融量化平台的股吧论坛用户评论数据、雪球网股民社交网络数据和财经新闻数据，分别构建股吧论坛投资者关注度、雪球网投资者关注度、新闻关注度和新闻情绪指数，并基于相关性分析和VAＲ模型探究这些社交媒体指标与沪深 300 指数的关系分析这些不同来源的社交媒体指标对股市影响的差异性。研究结果表明，不同来源的投资者关注和投资者情绪对中国股票市场的影响是不同的。在散户比例较高的中国股票市场，代表个人投资者关注的股吧论坛投资者关注度和雪球网投资者关注度对股市的影响是大于代表机构投资者关注和看涨看跌情绪的新闻关注度和新闻情绪指数的。其中股吧论坛投资者关注度对股市的影响最大，新闻情绪指数和市的关系不大。并且个人投资者关注的变动领先于股价的变动，股价的变动领先于新闻关注和新闻情绪指数的变动。部慧,解峥,李佳鸿,吴俊杰（2018）基于东方财富网股吧帖文与朴素贝叶斯方法,提出融合股评看涨看跌预期和投资者关注程度的投资者情绪度量指标。进一步,利用Granger因果检验、瞬时Granger因果检验、跨期回归分析等方法,探讨了投资者情绪对我国股票收益率的影响。实证结果揭示：月度和周度的股票收益率对投资者情绪的影响为负，而日度股票收益率对投资者情绪的影响为正。投资者情绪对股票价格存在当期正向影响，说明从市场整体来看市场的投资者情绪越高，对股票的需求越高，股价也就越高，反之亦然。石善冲等人（2018）以基于微信文本挖掘的投资者情绪与上证指数收盘价、成交量为研究对象，研究了投资者情绪时间序列与收盘价、成交量时间序列之间的关系。研究结果验证了投资者三种情绪倾向对股票市场的影响方式和效果不同：基于微信文本挖掘的投资者消极情绪比例能够稳定预测上证指数收盘价，基于微信文本挖掘的投资者积极情绪倾向和中性情绪倾向比例的增减变动能够迅速引发滞后１天的上证指数成交量的增减变动。研究表明基于微信文本挖掘的投资者情绪对于预测股票市场表现有重要作用。戴德宝等人（2019）通过文本挖掘技术和情感分析方法生成积极和消极各三阶共六类投资者情绪时间序列数据，采用单位根检验、Granger 因果关系检验和因子分析等方法构建上证投资者情绪综合指数，并分别使用支持矢量机和神经网络预测股票市场价格变化，进行假设验证。结果表明:利用网络股市论坛文本数据和股票交易数据构建的上证投资者情绪综合指数能够提高股指走势预测的精度，有利于政府、在线平台、上市公司和投资主体更好决策。

仍有少数文章已分析网络社群聆听捕捉投资人对个别股票的情绪信息。例如， Nguyue,Shirai and Velcin(2015)从雅虎财经留言板中以日頻率收集了一年共18只股票的留言板，从留言板的文本中提取出网络情绪。他們同時建立了一个利用社交媒体情绪来预测未来股票价格走势的模型，通过监督机器学习来训练该模型，并将投资人对公司特定主题的情绪纳入股票预测模型中。其实证结果发现，使用该模型对股票价格进行预测的准确性比仅使用历史价格的模型要高出2.07%的收益。Avery, Chevalier and Zeckhauser, (2016）研究大约250万个由个人用户提交到由Motley Fool company运营的“资本支持计划”网站的数据，并基于正向选股和负向选股的投资组合来预测未来个股股价的上涨和下跌。其结果证实卖空网站上投资人负面选股数量较多的股票和买入网站上投资人正面选股数量较多的股票的策略，在样本期内每年产生超过9%的回报率。

孙鲲鹏和肖星（2018）利用中国Ａ股上市公司的股吧发帖数据，用发帖情绪衡量投资者情绪，用帖子阅读量和跟帖评论量衡量情绪传染程度，探究了互联网社交媒体在投资者情绪传染中的作用以及对股价崩盘风险的影响。结果表明：投资者的发帖情绪越乐观，未来股价崩盘风险越高，这种现象源自乐观情绪在股吧的传播扩散，而机构投资者持股会加剧这种效应；社交媒体等互联网技术的发展会助长投资者情绪的蔓延并引起股价崩盘。段红娇等人（2017）选取东方财富网股吧论坛的个股帖子，使用计算机文本处理技术提取帖子情绪，结合证券分析师对个股的“中性”评级数据，实证研究了我国股票网络论坛的信息含量问题。研究发现:股票当日收益率受当日论坛情绪影响，为显著正相关;股票未来两日收益率与帖子数显著负相关; 股票当日的帖子数显著正向影响当日股价波动，而且能影响未来两日的股价波动;当日情绪分歧度越大，未来两日的交易量越大。

綜上所述，第一、孙鲲鹏和肖星（2018）沒有采用文本分析的方法。Nguyue et. al(2015)及Avery et. al (2016）文章都仅采用文本分析的内容，但其未采用文本分析的方法来处理。然而使用文本分析的方法赋予每个词频一定的权重，有 的优点，能更加精确地衡量了投资人的情绪 ()。第二、何學者??? 段红娇等人（2017）在构建指数的过程中只考虑到对个股看涨或者看跌情绪出现的频率。本文在构建情绪变量时，不仅考虑到了投资人情绪出现的频率，而且还赋予词频一定的权重。通过词频和权重，我们可以得到投资人看涨和看跌情绪的情绪密度，能够更精确度量投资人情绪。第三，除了段红娇等人（2017）分析投资人情绪和个股报酬风险之间的关系之外，其他文章都仅分析投资人情绪对股票报酬的影响。本文在此基础之上还研究投资人情绪对个股报酬风险的影响。综合分析个股的报酬和风险可以使投资人的投资行为更加理性，选股更加准确。但是，段江娇等人（2017）是通过分析投资人情绪的一致性而不是情绪的大小来探讨投资人情绪对个股报酬的风险的影响。第四、除了仅Avery et al. (2016)分析投资人对个股乐观及悲观分别的情绪之外，其他文章都仅分析投资人对个股乐观与悲观混合的整体情绪对股票报酬的影响。然而，分析投资人对个股乐观及悲观分别的情绪可更精确衡量投资人情绪对股票报酬的影响。第四、Nguyen，Shirai和 Velcin（2015）利用大型企业的股票数据预测了未来的股票报酬，但这些企业多属于于高新产业中的领头企业，在行业范围和公司规模上具有单一性。然而，本文样本在产业上具有多元性，故进一步提升了研究的普遍适用性。

然而，这些文献的投资人情绪信息来源多来自单一网络平台，可能较会捕捉到特定投资客群的投资人情绪信息。也就是说一些基础的金融科技技术虽可于特定平台大量的挖掘信息，但尚需要较成熟的金融科技技术(如网络爬虫技术)等才能广泛于各个网络平台搜集不同投资客群的投资人情绪信息，也才能结合较全面性的网络社群聆听获得较客观多源的网络声量。

本课题通过网络爬虫技术爬取各网络社群平台，直接捕捉投资人对某只股票涨跌看法的信息，形成乐观及悲观情绪的网络声量，并将正负面情绪的强度勾稽至金融大数据平台，探讨正负面情绪强度对股价报酬的影响进而预测报酬。

Avery, C.N. , Chevalier, J.A. , Zeckhauser, R.J. ,2016. The “CAPS”prediction system and stock market returns. Rev. Finance 20 (4), 1363–1381 .

Moat, H.S. , Curme, C. , Avakian, A. , Kenett, D.Y. , Stanley, H.E. , Preis, T. , 2013. Quanti- fying wikipedia usage patterns before stock market moves. Sci. Rep. 3 .

[1]杨欣,吕本富.突发事件、投资者关注与股市波动——来自网络搜索数据的经验证据[J].经济管理,2014,36(02):147-158.

[2]唐涛.基于大数据的网络舆情分析方法研究[J].现代情报,2014,34(03):3-6+11.

[3]李金海,何有世,熊强.基于大数据技术的网络舆情文本挖掘研究[J].情报杂志,2014,33(10):1-6+13.

[4]孟雪井,孟祥兰,胡杨洋.基于文本挖掘和百度指数的投资者情绪指数研究[J].宏观经济研究,2016(01):144-153.

[5]王靖一,黄益平.金融科技媒体情绪的刻画与对网贷市场的影响[J].经济学(季刊),2018,17(04):1623-1650.

[6]王夫乐,王相悦.社会情绪是否会影响股市收益——来自新浪微博的证据[J].山西财经大学学报,2017,39(02):35-46.

1. 段江娇,刘红忠,曾剑平.中国股票网络论坛的信息含量分析[J].金融研究,2017(10):178-192.
2. 石勇,唐静,郭琨.社交媒体投资者关注、投资者情绪对中国股票市场的影响[J].中央财经大学学报,2017(07):45-53.
3. 孙鲲鹏,肖星.互联网社交媒体对投资者情绪传染与股价崩盘风险的影响机制[J].技术经济,2018,37(06):93-102.
4. 石善冲,朱颖楠,赵志刚,康凯立,熊熊.基于微信文本挖掘的投资者情绪与股票市场表现[J].系统工程理论与实践,2018,38(06):1404-1412.

戴德宝,兰玉森,范体军,赵敏.基于文本挖掘和机器学习的股指预测与决策研究[J].中国软科学,2019(04):166-175.

**3. 研究内容**

**3.1 本课题的研究对象**

本计划以上海及深圳股市投资人为研究对象，搜集近十年来这些投资人每日在这两个股市所有相关社群网站对分别ETF个股发表的风险偏好、预期报酬、观点及看法的充分信息，纪录日期、内容、个股、关键词，并由内容及关键词初步筛选为乐观或悲观情绪。

**3.2本课题的总体框架**

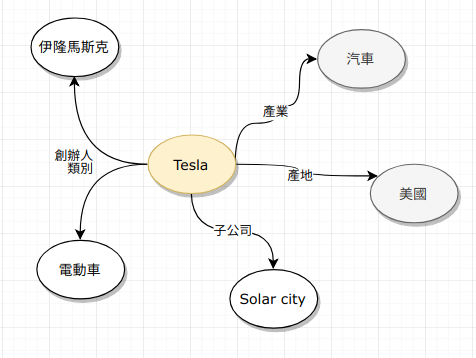
当影响个股公司或股市大盘的事件发生时(图1)，网络各社群平台讨论信息产生，本课题透过网络爬虫程序密集采集投资人对个股涨跌看法，结合社群聆听以语意解析来解读投资大众对公司发展及公司股价涨跌的评价。并采用「文本情感分析」来建构正面及负面情绪的网络声量；设立投资人个股情绪的数据库，并利用其来构建乐观及悲观情绪变量(图2)。进而将情绪变量勾稽金融大数据平台(图3)以建置模型，分析乐观及悲观情绪对股票报酬影响及对后续股票报酬做预测。本计划将事件讯息图谱、投资人情绪图谱、金融大数据平台运作图谱结合为知识图谱。以下绘图说明之：



1. **事件信息图谱。**主要提供机器人对于现实世界当中对于金融相关事件的理解。

新闻摘录：中国将对美国 600 亿美元商品课征 5~25%的关税，商品包含太阳能电池、汽车零组件、家电商品等，去年美国出口到中国的商品总金额为 1,300 亿美元。

可发现对于Tesla将会有重大的影响，因其属于汽车产业、美国公司、且子公司的业务亦为太阳能电池相关。



**(2) 投资人情绪图谱。**以下为情感分析范例：

|  |  |
| --- | --- |
| 评论 | 正负 |
| 此事件将使得Tesla公司受到严重冲击阿！ | -0.9 |
| 不可能真的课这么重的关税，美国政府知道严重性，过一阵子又会变卦了 | 0.6 |
| 这事件对电动汽车产业，尤其是对Tesla，或多或少都一些不利的影响 | -0.1 |

**(3) 金融大数据平台运作图谱。**将投资人情绪图谱所得到的正面及负面情绪汇入金融大数据平台，并予对应的相关事件所牵连的产业及个股进行交叉分析，探讨其对产业或个股股价及交易量的冲击程度。



**2.3 拟突破的重点和难点**

 (1) 透过网络爬虫的程序来密集采集投资人对个股乐观及悲观情绪相关的信息，需要一定人力、时间及经费投入来建置金融大数据平台，且也需要运用知识图谱来串接网络爬虫程序。(2) 语意解析初步分辨正面及负面情绪的语意数据须查询相关机构所建置的语意大数据数据库，须投入相对的人力及时间。(3) 本计划执行方式与过去的「量化投资」之主要差别在于运用机器学习技术，结合知识图谱来搜集投资人对个股乐观及悲观情绪信息，并串联事件发生对股价影响，为量化投资朝更精细化之技术。

**3.4 本课题的主要目标**

本计划的研究目标想透过网络爬虫的技术结合股市各投资客群的社群聆听，勾稽投资人语意解析与财金讯息，建构投资人对个股正面及负面情绪的网络声量，并建置模型分析乐观及悲观情绪对股票报酬率影响及对股票报酬率做预测。

4. **[思路方法]** 本课题研究的基本思路、具体研究方法、研究计划及其可行性等。

**4.1 研究的基本思路和具体研究方法**

**「第一阶段」：资料搜集**

语意解析引擎的建置需要大量的文本数据作为训练准确度之基底，故前期需要藉由爬虫的程序来进行网络理财投资的文本资料搜集的动作。本计划以上海及深圳股市自2009年1 月1 日至2018年12月底之10年期间日数据为样本期间；搜集数据报括每日投资人在这两个股市所有相关社群网站对分别上市ETF个股发表的风险偏好、预期报酬、观点及看法的充分信息，建置完整的股票投资人个股情绪的数据库，并定期更新。

**「第二阶段」：语意解析下结合网络社群聆听**

透过语意解析引擎立即进行语意的判断并做出反馈，而语意解析引擎的建构须结合知识图谱的建置，建构投资人语意解析与财金讯息勾稽之知识图谱将会布署在云端等适合的环境中，最后再结合不同投资客群的社群聆听来解读投资大众对未来公司发展及公司股价涨跌直接与真实的评价，来与实际股价走势来相互印证。

**「第三阶段」：划分乐观及悲观情绪的术语**

本计划分别针对投资人乐观及悲观情绪的*lth*类别来搜集所有潜在的特征词，并计算*lth*类别的*jth* 特征词的词频。如 *rO*和*rP* 分别是乐观及悲观情绪的词频。在和乐观及悲观情绪相关的许多特征词中, 代表性特征词相较于其他者更攸关。我使用卡方检测以决定代表性特征词(Yang and Pedersen, 1997)。首先划分属于乐观、悲观及其他(和前两类无关)类的特征词。检定每个特征词是否属于特定类情绪，每次鉴定两类(*l*=2), 如乐观及非乐观。虚无假设是*jth* 特征词属于非乐观类，而对立假设是 *jth* 特征词属于乐观类。悲观类可设计为相似的假设作法。卡方独立性检定的统计量为下列公式:

 (1),

此处A和B是指在*l*th类*j*th和non-*j*th 特征词的词频; C和D是指在non-*l*th 类*j*th和non-*j*th 特征词的词频;且N是指所有特征词的词频。假设*j*th 特征词独立于*l*th类。若卡方值高于临界值，独立于乐观的虚无假设被拒绝；故判定*j*th特征词被视为乐观(*l*th)类的代表性术语。

**「第四阶段」：估计情绪特征词的权重**

本计划接续采用焗权重法分别计算乐观及悲观情绪的特征词之权重。透过权重可指出特征词的强度。此方法已被广泛使用在语意分析做为所有代表性特征词相对大小的衡量。以乐观情绪为例，焗值公式如下述:

 (2)

此处*m* 是厂商的数目; 是投资人对*i*th 厂商乐观情绪中*j*th 特征词的词频; 且 是乐观情绪中*j*th 特征词的焗值。本研究接续分别计算乐观及悲观情绪的*j*th特征词的权重。

 (3)

 (4)

此处n 是投资人乐观或悲观情绪特征词的数目。权重愈大表示乐观或悲观的特征词愈重要。

**「第五阶段」：量化产生乐观及悲观情绪变量**

此阶段计算投资人在某月*t*对*ith* 厂商的“乐观情绪密度”(*IOSit*)及“悲观情绪密度”(*IPSit*)。藉由对乐观情绪计算特征词，量化投资人对特定厂商的乐观情绪之程度。本计划首先将投资人每日对*i*th厂商乐观情绪 *j*th特征词的词频乘以相对应的权重，并累积对乐观情绪的所有代表性特征词，计算为每日*IOS* (*IOSidk*)。接续，我加总每日的*IOS*计算为每月的*IOS*, 并加总每日的*IPS*计算为每月的*IPS*.

 (5-1)

 (5-2)

此处下标*i, d,* 和 *k* 是指*ith* 厂商在*dth*天*kth*则网络信息; 是指投资人对*ith* 厂商在*dth*天*kth*则网络信息为乐观情绪的*j*th 特征词的词频；则是指相对应悲观情绪的词频；且和是投资人乐观及悲观情绪的*j*th 特征词的权重。被计算的*IOSit*和*IPSit*可随时追踪及更新投资人对*i*th厂商乐观及悲观情绪的程度。因为*K* (网络信息的数目) 和*n* (特征词的数目) 是无限的, 这两个密度变量的最大值是无限的，本计划分别对*IOSit*和*IPSit*将非0的密度值，使用均等分配的公式[99\*(*x* – *x*min)/(*x*max-*x*min) +1](*x*是密度值)，将之标准化介于最小1% 和 最大100%之间；然而0密度值仍然为0。较高的密度值意谓投资人有较强的乐观或悲观情绪。

**「第六阶段」：设立回归模型─检测情绪对股票报酬的影响及预测**

最后本计划使用下列的OLS回归，以检测投资人对厂商的乐观及悲观情绪对此月及后续月该股票累积异常报酬的影响：

 (6)

此处 *i* 是在本样本*i*th厂商, 且*t*是*t*th月。*CAR* 是股票*i*于*t*至*t+j*月(*j*=0, 1, 2, 3)的累积异常报酬, 此处异常报酬AR是从市场模型所估计。估计期涵盖在检定期之前250个交易日。本计划认为投资人情绪影响股票报酬应该不会超过一季。*AR* 样本期间是从2009年1月第一个交易日至2018年12月底最后一个交易日。*IOS*是指投资人对厂商乐观情绪的程度, 而*IPS*是指投资人对厂商悲观情绪的程度。是横断面股票报酬的控制变量，包括*Beta*, *Size*, *BM* 和 *Momentum* (Fama and French, 1992; Jegadeesh and Titman, 1993)。每股 “*Beta*”是由估计市场模式得到，和*CAR*作法相似。*Size* 是每个厂商流通在外股票的去年底市值的对数值。*BM* 是指净值市价比。*Momentum* 是指每股报酬减检定期前一年的报酬。最后采用真实报酬数据进行回溯测试，检测以此模型预测报酬的准确程度。

Doms, M., and N. Morin, 2004, Consumer Sentiment, the Economy, and the News Media, working paper, Federal Reserve Bank of San Francisco.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. J. Econ. Perspect. 21, 129–157.

Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62 (3), 1139–1168.

Da, Z., Engelberg, J., Gao, P., 2015. The sum of all FEARS investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32.