西南财经大学

SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

MASTER′S DISSERTATION

上市公司新闻对市场股价影响
——基于 TRANSFORMER 框架的研究

The Impact of Listed Company News on Stock Price
-- A Study Based on TRANSFORMER Framework

学位申请人 韩逸凡

指导教师 郑羽

学科专业 金融工程

学位类别 经济学硕士

分类号 密级

U.D.C

上市公司新闻对市场股价影响
——基于 TRANSFORMER 框架的研究

The Impact of Listed Company News on Stock Price
-- A Study Based on TRANSFORMER Framework

学位申请人：韩逸凡

学　　号：220020Z2001

学 科 专 业：金融工程

研 究 方 向：金融科技

指 导 教 师：郑羽

定 稿 时 间：2023.02.15

摘要

在资产定价的研究中，市场上存在以文本为主的大量的非结构性数据，蕴含了大量价量数据所未包含的信息。但想要准确挖掘这类信息，往往需要借助复杂的人工智能算法以及大量标记数据，使得市场研究与投资分析受挫。

本文针对金融市场中的新闻文本数据，区别于传统的机器学习方法，引入了 Transformer 框架，在中文语料库上构建了自己的金融 BERT 大型预训练语言模型，并且通过多组对比实验，在平衡试验成本与模型精度的基础上，尝试构建了用自然语言处理大模型解决金融问题的一个初步方案与流程。随后，研究使用最优模型，对 A 股市场上 170 万条新闻进行分析，结合因子定价基本方法构造新的情绪因子，发现通过这种金融科技结合方案能得到较好的实证效果。

通过对 A 股市场上带有新闻的个股进行投资组合的划分并回归，我们验证了金融媒体情绪的确会影响市场投资组合收益率这一实证结果，发现媒体情绪越消极的投资组合，其预期收益率越大；个股规模越小的投资组合，其受金融媒体情绪的影响越大。此外，我们还在沪市与深市，以及五个行业中分别进行了实证检验，证实了结果较为稳健，金融媒体的情绪因子隐含着关于市场的定价信息。

关键词：BERT Transformer 因子定价 自然语言处理 媒体情绪

ABSTRACT

In financial markets, there is a large amount of unstructured data, mainly in the form of text data, which contains a wealth of information not covered by price and volume data. However, to accurately mine such information often requires the use of sophisticated artificial intelligence algorithms and large amounts of labelled data.

This paper focuses on news text data in financial markets and, introduces the Transformer framework. Combined with self-supervised learning techniques, we build our own large-scale financial BERT pre-trained language model on a Chinese corpus. Through multiple comparison experiments, we attempt to create a preliminary solution and workflow for solving financial problems using large natural language processing models. Ultimately, this study uses this solution and model to analyze 1.7 million news articles in the A-share market. By integrating traditional financial paradigms, we construct new sentiment factors and find that this combination of financial technology yields good empirical results.

By dividing and regressing investment portfolios with stocks accompanied by news in the A-share market, we verified the empirical result that financial media sentiment does indeed affect market portfolio returns. We also found that the more negative the media sentiment in an investment portfolio, the greater its expected return. Smaller individual stock portfolios are more affected by financial media sentiment. In addition, we conducted empirical tests separately in the Shanghai and Shenzhen markets and five different industries, finding our results to be quite robust. The sentiment factors of financial news media imply pricing information about the market.

Key Words: BERT, Transformer, Media Sentiment, NLP, Factor Pricing

目　录

1.引言

1.1 研究背景

价值投资一直是一个被投资者们津津乐道的热门话题，在国内与国际证券市场上，有不少专业投资者、股票分析师以及研究学者，不懈挖掘市场内潜藏的信息，以此解释投资股票获得超额收益的原因。通常，分析师们使用公司的财务报表、市场收益率以及专业评级等结构性数据，对公司、行业以及整个资本市场的表现进行预估，进行资产配置以及风险评估。在金融及其衍生产品市场上，谁能够越快、越多得获得定价信息，解释收益产生的原因，谁就可能获取更大的回报。但以马科维兹的证券组合投资理论作为证券市场定量研究领域的开端，在 Fama&French（1993）提出三因子模型的基础上，越来越多的研究者不断挖掘股票定价因子，因子动物园的规模也愈发庞大，相应的，传统结构性数据所蕴含的潜在信息也越来越少，大量因子相继失效。

但是，伴随着社会生态不断向信息化、数字化迈进，大数据日常生活与经济发展中占据的比例越来越大，其重要性也越来越强。根据中国信通院的统计报告显示，截止 2021 年，全球数据总量已经高达 67ZB，其中我国数据总量占全球数据总量的 9.9%，高达 6.6ZB，而图像、文本等非结构性数据占已经占据了数据总量的超 80%。另一方面，得益于诸如云原生、分布式集群以及 AI 工程化等数字化技术的蓬勃发展，大模型、多模态的复杂高精度神经网络等人工智能模型的训练、测试、优化与部署已经变得愈发方便与简洁，由此在处理大数据、异类数据以及复杂研究场景的问题上取得了令人瞩目的发展。不少分析师与专业投资机构逐渐将目光转向新闻、研报等文本类非结构化数据。2019 年 Morgan Stanley 曾发表研报《The Power of Words: Going Global》，率先使用了深度学习模型对金融文本进行建模，挖

掘出文本数据中潜在的定价因子。Morgan Stanley 以自身机构对全球市场股票的 60000 份研究报告为数据，通过构造卷积神经网络计算出[-100,100]的情感得分，将其作为情感因子在全球市场进行测试，得到了较高的回报，并且在规律在美国、欧洲、亚洲市场都得到了证实。国内机构中，华泰金工陆续发布了 40 多篇人工智能研报，其中 2 篇阐述了从 Wind 新闻数据中构造情感因子，并进行回测的过程，尽管构造方式不同，但该因子同样的到了不错的收益。

图 1-1　MORGAN STANLEY 研报情绪因子收益

同时，不仅业界在想方设法挖掘文本数据中的潜在信息，力争获取高额回报，近些年来越来越多的学术研究也将目光投向了与市场有关的文本信息上。一个显然的事实是，多篇研究指出金融媒体的情绪能够一定程度的影响资本市场的表现，特别是近年来，诸如纸质媒体这类的传统媒体逐渐退出市场，同一时间网络媒体包括自媒体与流媒体日渐兴盛的情况下，其对资本市场的影响力愈发凸显。例如在 2015 年 4 月，有不少媒体纷纷在新闻报道中提出看法，即“4000 点才是 A 股牛市的开端”，这一观点无疑增强了市场中部分券商分析师对市场的信心，从而导致整体信心指数环比增长了 21%左右，但随之而来的却是 A 股市场上一次触目惊心的股灾。与此相反，媒体情绪对市场的一次负面影响事件发生在 2019 年底，白酒行业被曝出添加甜蜜素的事件，各路媒体争相报道，负面舆情爆发，最终白酒板块遭受了重大负向冲击，整体股价大幅下降。

这或许是因为，媒体在叙述报道基础事实的基础上，或有意或无意都会向读者输出一些具有倾向性的观点，这其中包括了对公司现今市场表现与经营业绩的评价，以及对公司未来发展的一些展望。毫无疑问，这些评价与预期很难保持一个完全中立的态度从而丝毫不影响投资者、分析师以及资本市场，这表明媒体情绪在资本市场上是客观存在的。Boudoukh et al.在2019年的研究也表明，媒体舆情会引导市场注意力和交易行为，极大地影响了资本市场的价格波动性特征，这同时也印证了Kogan et al.在2018年的研究成果，这甚至会引起股票价格过度反应与反转效应(郦金梁等，2018)。因此，媒体不仅仅是传统认知上中立和客观的信息传播者，而且在报道新闻时也会向公众传达自己的立场和态度，甚至去引导舆论倾向或是干预市场情绪。通常来讲，无论是官方媒体还是自媒体，通常都在传播学领域充当意见领袖的角色，媒体的情绪往往会在公共环境中迅速而广泛地传播，甚至被过度重复和强调，使其在投资者群体中交叉传播（游家兴、吴晶，2012）。这种公共空间的舆论环境会广泛影响资本市场的投资者对股票、指数甚至其他衍生产品走势或公司业绩以及公司预期的判断和认知，从而影响到投资者与分析师的理性情绪，甚至是金融资产的理性定价和准确预测发生偏离(Tetlock et al.，2008; Fang and Peress，2009)。同时，据不完全统计，国内市场上约80%的投资者为个人投资者，与专业机构培养的投资者不同，他们往往更加情绪化，缺少理性投资的经验，从而受市场上媒体情绪的影响可能更为显著。

那么，如何有效、快速、准确的提取媒体情绪信息，并将其转化为证券市场上的定价因子，将会是一个值得研究的话题。另一方面文本等数据虽然有大量潜在信息亟待挖掘，但是不同于类似时间序列等结构性数据，新闻文本难以直接运用于统计学模型并在金融市场上进行实证分析。传统的文本分析方法多采用词频统计以及相应的衍生算法，割裂了文字间本身的上下文关系，导致预测结果严重依赖于主观经验与统计手段，不仅筛选出的模型变量存在一定的滞后甚

至变量遗漏的现象，而且在情感理解和准确度把握上容易出现失衡，从而导致结果偏差严重。得益于近些年来自然语言处理（NLP）技术的飞速发展，如何高效处理文本数据，将抽象的文字做数字化并进行分析与建模的难题得到了充分的解决。但往往为了训练出高精度的模型，得到较好的分析效果，不仅需要极其复杂的模型结构，也需要大量数据进行训练与验证，并且数据本身要求质量极高，数据中所蕴含的专家知识的好坏对模型的优劣也有着十分显著的影响。此外，高精度的模型，往往也对训练、部署需要的机器有着极高的要求，不仅要求一定程度的运行速度，其巨大的数据量也决定了机器需要大量的内存要求。这些因素也成为了金融科技研究与量化投资领域中使用人工智能模型分析非结构性数据的最大门槛与障碍，导致类似的分析往往被大机构、大公司所垄断。

本文着眼于金融媒体情绪对资本市场上股价及其收益率的影响，为了增强对金融新闻文本分析的精度，引入了基于 Transformer 框架的 BERT 模型（Bidirectional Encoder Representation from Transformers），在使用少量的标注数据的情况下，对金融新闻文本的情绪分析得到了较高的预测精度。并依据该模型对市场上的金融新闻进行情绪分类，构造情绪因子，并参照 Fama-French 的因子模型，在 A 股市场上进行了充分实证，验证了其具有定价能力，能够一定程度的解释市场的超额收益。并且，本文还在 A 股不同市场以及不同不同行业分类中验证了该因子的定价能力，使得结论具有一定的稳健性。

1.2 研究意义与创新

本文主要有两个研究意义。第一，现有的多因子模型中，仍以使用结构化数据构造因子的研究占据绝大多数，对于市场上情绪的分析大多从投资者入手，例如构造换手率因子。仅有少数基于统计词频或者 SVM 等机器学习方法从文本入手，分析媒体或者投资者的情绪因素，并且也少有从因子定价的角度论证金融媒体情绪对市场影响

的研究。本文将基于多因子定价的方法，使用人工智能的算法引入情绪指数构造媒体情绪因子，结合中国版三因子模型，探究媒体情绪因子对上市公司股价的影响。

第二，区别于传统的统计学习方法，依靠神经网络强大分类能力的深度学习算法构建起来非常复杂，并且为了得到较好的模型效果，往往需要海量数据参与模型训练，这无疑增大了绝大部分学者、投资者与分析师的研究门槛。例如，JP Morgan 使用 CNN 构造的情绪指数来选择投资组合，其模型本身并不复杂，但成本最高的是其内部的专家对训练所使用的的 60000 份专业研报的评判与标记，这将极大程度的影响最终对金融文本的分析效果。另一方面如果降低专业数据门槛并简化模型结构，模型对文本的分析能力也将大打折扣。因此，本文尝试将目前深度学习中最热门自监督学习 SSL 框架与最强大的网络模型 Transformer 进行结合，在使用较小样本的基础上，探究能否生成对金融文本有着较好特征挖掘效果的网络模型。

由此，本文的创新点主要在于，一是通过引入前沿算法，并充分对比模型，希望能为后续金融科技相关研究在模型选择上提供参考；二是通过挖掘出的情绪信息构造新因子，希望能为监管部门或投资者理性看待市场上的媒体新闻提供建议；最终，本文希望通过这种学科交叉的视角，能在一定程度上拓展因子定价理论研究的技术路线。

1.3 研究方法

本文遵从从理论到实践的研究路线，先针对问题以及相应领域梳理大量文献，厘清该问题的研究脉络、研究途径以及重要结论与成果，并根据本文研究内容的具体情况，选择合适的方法与严谨的数据对研究所提出的假设进行检验，期望能在实验中得到结论的证实。

理论方面，本文将从人工智能算法与金融因子定价实证两个方面出发。一方面，深入研究深度学习自然语言处理领域中的模型原理与模型搭建细节，为后续的模型训练与部署奠定基础。另一方面，本

文从基本的三因子模型出发，探究如何引入新的因子，并对其因子定价能力进行实证分析与检验。

模型搭建方面，本文将依照现有的自然语言处理大型预训练语言模型的框架，在金融文本语料上进行充分的预训练，并尝试在下游接入多种分类器，与现有开源的模型进行对比。选取其中效果最好、部署成本相对较低的模型对资本市场上的媒体情绪进行分类。

实证方面，本文首先探究如何使用模型生成的媒体情绪分类指标合理构造媒体情绪因子，并进行基础的因子统计性描述。在此基础上，本文将构建金融媒体情绪的因子，并结合中国版的三因子模型，考察其在 A 股市场上的定价能力。并且，我们还将与中国版三因子模型进行对比，查看引入的新因子是否有新增信息增量能解释额外的超额收益；最终，本文还将从多个市场、多个行业考察情绪因子的定价能力，检验其稳健性。

1.4 研究框架

本文第一章节为绪论，主要阐述本次研究的金融领域现象背景、原因，在机器学习与金融实证结合方面的研究的意义与主要贡献，以及论文研究的方法和思路。该章节着重介绍了以文本数据为主的非结构化数据在金融市场投资与研究的重要性、当前分析金融文本时平衡精度与研究成本之间的难点以及本文打算采取的解决办法，并介绍了研究的内容与框架。

本文第二章节为文献综述，以国内外相应研究内容与研究脉络为载体，分别阐述了金融市场上因子定价的重要意义与主要问题，文本数据在金融研究中的重要性以及先有金融领域文本研究的主要方法与重要结论，最终简述了自然语言处理研究与技术的发展历程，为解决金融领域的媒体情感分析与相关因子构建提供思路。最后将上述三个领域进行总结，为本文的研究指出理论方向。

本文第三章节为 NLP 预训练语言模型的搭载、训练、测试与部署，主要结合了谷歌团队研发的双向转换器词编码模型（BERT）与中

文词掩码预训练任务（Whole Word Mask），在金融新闻语料上进行预训练。并使用少量标注数据在下游进行新闻情感分类任务的微调（Finetune），并最终在多个模型上测试效果，为后续的因子构建与实证分析提供可靠的模型支撑。

本文第四章节为媒体情绪因子的构建以及其定价能力的验证，观测新因子的基本统计性质以及收益能力，并使用 Fama-Macbeth 回归的方式，对多因子模型进行定价能力以及因子影响方式的实证分析，观测其模型系数的显著性，论证因子的定价能力，并从多个市场以及多个行业检验该因子的稳健性。这一章节将得出关于媒体情绪因子定价有效性的实证证明。

本文第五章为研究的主要结论，主要从金融科技的结合应用与金融媒体情绪定价两个方面展开总结，梳理了本文在引入人工智能大模型辅助金融实证研究方面的主要工作与成果。在此基础上简要分析了后续使用金融媒体情绪因子在 A 股市场上探究资产定价的市政结论，以此为案例，可以一定程度上说明金融科技交叉领域结合的可行性与实用性。除此之外，在本章节，我们还并提出了研究现有的不足之处，主要是针对新闻文本的预处理不够深入，可能还有不少潜在信息值得挖掘，可以结合图神经网络、多模态模型来丰富解决，这一部分为以后的深入研究指出了方向。

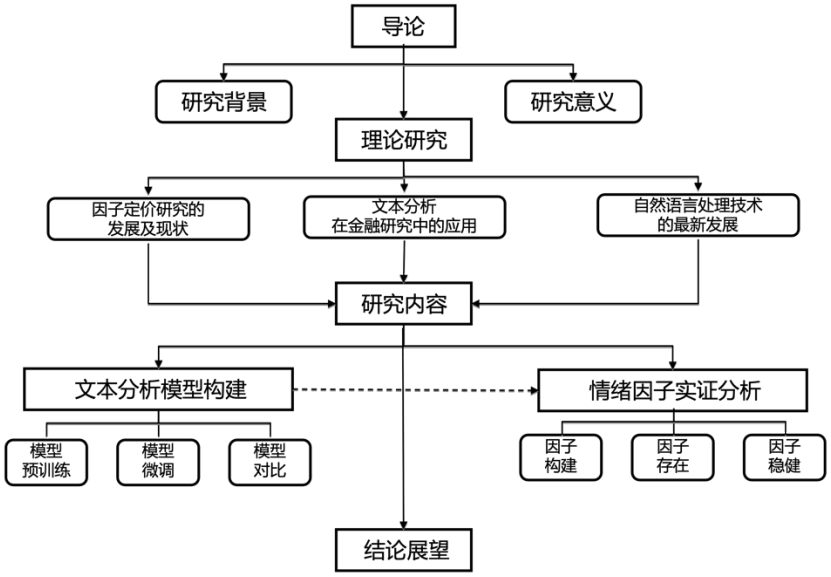

图 1-2　整体研究框架与内容

2.理论分析与研究假设

2.1 资本市场与因子定价

在现代的社会经济活动中，资本市场能够为企业提供筹措资金、加速融通的场所，还能够帮助公司完善其治理结构。不仅如此，它还能够帮助投资者实现自身的资源配置，也是实现社会资源配置的重要平台。而现代投资学的精髓，最本质的特点体现在如何把控投资的风险，这使得投资者们将研究问题聚焦在两个方面，即如何平衡“预期收益最大化”与“风险最小化”两大目标。投资者的投资组合问题到底是什么呢？在 20 世纪 50 年代，马克维茨(H. Markowitz)认为，其应该投资者是为了解决怎样在不确定的条件下将自己的投资效用最大化，也就是如何在风险可控的情况下，尽量获取足够大的收益。因此，他在论文中提出了著名的证券投资理论，这也为现代证券投资理论的发展奠定了基石，标志着现代证券理论研究进入了定量分析阶段。随着现代投资理论的深入发展，20 世纪 60 年代，美国学者 Sharpe(1964)、Lintner(1965)和 Mossin(1966)在马克维茨的证券组合投资理论和托宾的两基金分离定理基础的上，进一步研究了市场达到均衡时资产收益与风险之间的关系，得到了著名的资本资产定价模型(Capital Asset Pricing Model)，简称 CAPM，这也是第一个金融资产定价的均衡模型，它有效地揭示了投资者按着 CAPM 去做会得到什么样的投资结果。CAPM 还为投资者提供了一种有效的投资办法，投资者可以根据该模型研究所要投资证券的系统风险(β 值)，进而根据系统风险的大小确定合适的回报率去选择投资证券。

自从 CAPM 模型被提出以来，在金融学界引发了热烈的反响，早期的实证研究实验都发现股票的风险和收益可以拟合成一种正向的线性的，即当投资者会承担更大的风险时，相应的投资组合所得到的预期收益就应当更高。其中较为著名的是 1972 年，Black、Jensen

与 Scholes 使用 1926 到 1966 年纽约股票交易市场的股票数据对 CAPM 模型分别进行了时间序列和横截面的检验，为β系数和股票收益率之间存在着正向线性关系的结论打上了一剂强心针。随后在 1973 年，Fama 和 MacBeth 采用了相同的方法对 CAPM 进行检验，在实验中，区别与之前的做法，他们第一步使用时间序列的回归得到风险系数，以此用来预测投资组合未来的可能收益率，最终实验结果虽然同样说明了股票收益与β系数之间正向的线性关系成立，但他们却发现 CAPM 模型中的β系数对投资组合超额收益并不能进行很好的解释。

1993 年，Fama 和 French 使用美国市场上 1963-1990 年的数据，对股票收益率进行回归分析，观察到虽然 CAPM 中的β系数不能对收益率进行很好的解释。然而，他们发现，股票公司的市值和账面市值比率能更好地解释回报率。由此，Fama&French 通过对财务指标大小进行分割，将满足相应比例的股票构造成一个投资组合，从而计算出对应的因子收益率，Fama-French 三因子模型由此诞生。后来，Fama 和 French（1998）在不同国家的市场上，用三因子模型在不同的时间区间内进行实证研究。首先在 13 个主要的证券市场上，市政发现其中有 12 个市场印表现，能够验证账面市值比的对市场股价的解释能力；此外，在 16 个新兴市场上，Fama 和 French 也能够使用规模因子对超额收益率进行解释，从而发现了规模效应。实际上，因子也可以理解为是不可分散风险的来源，当暴露于某一个因子的风险下时，投资组合因为承担了该因子不可分散的风险，而获得更多的风险溢价。在 Fama-French 之后，金融市场上立刻掀起了一股因子挖掘的热潮，这其中最著名依然是 Fama-Frech（2015）提出的五因子模型，在三因子模型的基础上，引入了盈利因子（RMW）与投资因子（CMA），发现价值因子（HML）变得不再重要，其定价信息被其余因子所囊括，Fama 进一步研究认为，主要是盈利因子与投资因子，这两个新因子囊括了价值因子的绝大部分定价信息。在因子定价研究方面，学术界往往更加严谨的区分异象因子和定价因子，因子必须能

够 解释资产预期收益率截面上的差异。由此，学界往往从经济学和金融学的原理出发，从大量相关的因子中找到有限个相对独立的因子构成多因子模型，而把其他没有被加入模型中的、能获得模型无法解释的超额收益的因子都视为异象。

在国内因子研究方面，由于国内资本市场起步时间较晚等客观原因，其市场机制与市场结构暂时还没有发展的比较成熟，不仅由于非专业投资者较多而存在明显的投机行为，并且市场上许多公司的财务数据不能确保真实性、其相应的信息披露也进行的遮遮掩掩，这些问题依旧较为严重。有趣的是，与三因子模型在美国市场的稳定表现不同，国内不少进行因子研究的学者，在迁移三因子模型到国内做实证分析时，并未得到像 Fama-French 在美国市场那般取得显著效果，这可能也是其中原因之一。例如，黄兴旺和胡四修（2002）就确实发现账面市值比效应在中国市场中并不显著，利用 1995 年至 2000 年中国市场的股票数据进行实证分析，发现账面市值比因子并不能很好的解释市场的超额收益。但是，他们也发现 A 股市场上的规模效应却较为显著。同样的，关于账面市值比因子在国内 A 股市场的表现，顾娟和丁楹（2003）也和黄得出了近似的结论，最终认为中国市场上账面市值比效应基本不存在。但另一方面，也有学者在中国市场上对三因子模型进行实证，并对因子的定价能力得到了证明。例如范龙振和余世典（2002）发现 A 股市场股价具有显著的市值效应、账面市值比效应。后来也有学者尝试在 A 股市场上构建新的因子，并结合 Fama_French 三因子模型，构造更大的回归模型。潘莉和徐建国（2011）以沪深 A 股为研究样本，从而提取出了市值因子和市盈率因子，因其实证发现市值因子与市盈率因子对股票的收益率的影响较为显著，而加入杠杆率因子后，该因子对回报率的影响却呈现出不对称性，其在实证分析前期对解释 A 股超额收益有着明显作用，但在试验后期，其影响力逐渐降落。近似于黄兴旺和胡四修在 2002 年以及顾娟和丁楹在 2003 年发表结论，潘莉和徐建国（2011）发现在 A 股市场上，账面市值比与流通比率对股票收益并没有显著的影响。

Liu et al.(2018)认为，A 股市场存在着较为严重的“壳污染”现象，因此他们在剔除了 A 股市场市值最小的 30%股票的基础上，并使用市盈率因子（VMG）代替单因子模型中的账面市值比因子（HML），最终发现其在 A 股市场上对比传统三因子模型有着更好的解释力，并引入投资者情绪因子，在 10 个市场异象中，均对 A 股市场进行了测试研究。

如今，针对因子定价的研究方兴未艾，并且通过机器学习算法挖掘因子的方式也已经为许多人接受并使用，国外已有团队整理了 86 个通过机器学习算法挖掘出的因子，并且发现他们的定价能力不输传统因子，构造出的量化策略也能获得更高的回测收益。几乎同时，国内 Li et al.(2021)将这些因子在国内市场上做了验证并将其扩大到 94 个。目前，学术界已经挖掘出超过 400 个因子，整个因子数量以及因子定价的实证体系已经非常成熟，但是目前的主流因子几乎都还是从结构性数据中挖掘出来的，对于非结构数据中的信息，仍然需要研究进行实证探索与挖掘。

2.2 金融文本分析

在金融市场中，不同于图像、视频、语音等非结构性数据，文本大数据往往也与市场表现以及市场的一些行为存在千丝万缕的关系。在经济学领域,文本大数据能够被用来解决许多问题,例如 Baker et al.在 2016 年用统计学的方法梳理官方的政策文件，从而刻画经济政策的不确定性，而(Hoberg and Phillips, 2016)则是用文本对行业进行动态分类、Shapiro 与 Thorsrud 分别于 2018 和 2019 年，用官方的政策文件来度量社会经济发展并预测社会经济周期，此外，Gentzkow 和 Shapiro 也曾尝试用社会经济新闻来度量媒体报道偏差以及对社会的影响，从而指出社会对新闻的需求。而在金融学领域，现有的研究通常使用文本数据来度量政策等的不确定性(Baker et al.,2016)，刻画媒体或者社会活动参与人对某些是、某些领域的关注度 (Fang and Peress,2009; Daetal, 2011; Hillertetal,2014;

Ben-Rephaeletal ,2017)，亦或者通过抓取文本中的情绪或语调(Antweiler and Frank,2004)分析背后的社会经济现象，更有比较深入的研究，从新闻中提取隐含波动率，从而在市场上对证券、指数及其衍生产品进行定价，优化投资组合策略。

在这些金融领域的文本数据研究中，文本情绪正是现今的一个热点问题。在通常的新闻或者短讯中，金融市场的文字活动通常以积极与消极、看涨与看跌、做多与多空、牛市和熊市等方式来表达。而根据情绪来源的差异,我们也可以将文本情绪的研究对象分为媒体语调(即媒体新闻情绪)、管理层语调(上市公司年报的管理层讨论与分析、盈利电话会议和其他公开信息披露文件)、投资者情绪 (金融相关网络论坛发帖)等。

学界和业界通常都认为，证券分析师具有极高的职业素养，他们作为资本市场的重要参与者，负责深入分析市场信息并将可靠结果提供给其他投资者，极少收到金融市场上相关情绪的感染，是优秀的理性投资人。然而近年来，行为金融学领域的研究结果打破了这一经典假定，Easterwood 与 Nutt（1999）证实了证券分析师非大家认为的那样完全理性，而是与普通投资者一样存在认知偏差(Easterwood and Nutt，1999)亦或是心理偏差(Sedor，2002)，这将导致他们也会在市场上收到一些情绪的影响，例如，当他所服务的投资者产生情绪波动时，分析师的情绪难免也会受到一定感染(伍燕然等，2016)。而 DeLong et al（1990）认为，这种分析师与投资者情绪的波动，极有可能导致金融资产偏离正常的价格。

然而，不只在信息的接收端，新闻媒体，作为市场上权威信息的发布端，也是带有自身情绪波动与感染力的。在信息的发布、传播以及聚集的过程中，市场往往需要媒体充当一个发挥重要媒介功能的角色。然而事实上，媒体在信息的传播链条中，并不只是担任一个信息中介的角色。一些研究通过集中考察媒体对投资者行为、资产定价、公司治理与市场监督等维度的影响发现，媒体不光承担着“信息供给”的职责，他们也有可能做出“情绪干预”的行为。从媒体本身

的职责出发，其履行信息挖掘的重要职能(黄俊、郭照蕊，2014)的同时，就会对投资者的行为产生影响，甚至进一步影响到上市公司的决策行为以及治理效率(Dycketal.，2007)。不仅如此，一些媒体为了追求点击量的数值，追求更多的观众以及受众，其往往会选择一些比较偏激的角度进行理解，甚至与玩弄一些文字游戏，在报道时倾向于选择具有冲击性的表述(Hermida et al.，2012)；亦或者是在新闻撰写的过程当中，暗自藏匿传播其其情绪和观点，并通过文字进行输出，以此达到引导公众对某一热点事件的态度、看法甚至认知的目的(Mullainathan and Shleifer，2005)。随着越来越多的学者关注到新闻媒体存在的报道偏差与特有情绪等异象，媒体报道的客观性与媒体功能的有效性引发了广泛质疑(Rinallo and Basuroy，2009)，新闻媒体在提升资本市场信息效率与上市公司治理水平等方面所发挥的作用面临严峻拷问。

国外研究人员对媒体情绪的度量一般是从研究其与证券市场的关系开始的，通常使用《纽约时报》、《华尔街日报》以及《华盛顿邮宝》等权威数据来进行度量，进而分析这些情绪对整个交易市场以及相应的个股公司有什么影响。Tetlock（2007)收集了《华尔街日报》上的专栏文章，并统计了其中表述情感较为消极的词语频率，检验分析之后发现，当消极词语的次品上升是，股市的收益率会随之下降。除此之外，他还研究了负面词词频与公司股价收益率的关系，通过将《华尔街日报》、道琼斯新闻社上的数据与标普 500 指数相关联的公司进行匹配，得到约 35 万条数据，最终发现，负面词频越高的公司，不仅下一个交易日时股票的个股收益率会下降，下个季度时公司的盈利都将更低。此外，也有研究将目光放在情绪以及周期的非对称性上，Garcia（2013)就曾将经济的繁荣期与衰退期分开研究，在这两个时期内，他分别查看了金融新闻中蕴含的媒体情绪对资产价格影响，发现结果存在不对称性，即只有在衰退期时，新闻文本才能通过其中的消极词频的统计，获取在日度频率上对下一个交易日收益率的预测能力。此外，Garcia（2013）还考虑了正面、负面语调作用

的非对称性，研究发现，无论语调是正是负，其媒体情绪都能够预测大盘在下个交易日的收益率。而(Zhang et al;2016)却发现，在媒体报道出来的相关公司新闻中,正面和负面的词语比例对股票市场变量，如收益率、波动率、交易量等的影响是不对称的,因此在实证研究的过程中，建议应同时考虑文本中的正负语调。

而国内相关的媒体情绪的研究，虽然这些年也逐渐丰富，但受碍于起步时间较晚，无论是在媒体情绪考察和度量的方法上，还是媒体情绪与市场变量之间的实证研究上，几乎都还处于一个比较基础的阶段。从文本分析方法的维度来看，现有研究常用的主要有词汇词典法、词汇加权法、支持向量机以及 LDA 主题模型,这些模型在一定程度上也都有有不错的效果。游家兴和吴静（2012）选取了 2004-2010 年国内 8 家主流财经报纸上的新闻,在提取媒体情绪时，他们使用的是人工阅读的方式，一条一条从新闻中抽取相应的情绪指标，并结合这些情绪，实证分析对沪深 A 股市场的上司公司资产定价错误与媒体语调之间的关系；而汪昌云和武佳薇（2015）使用 6 家主流财经媒体的新闻数据,结合自己定义并构建的财经媒体新闻情绪词典，对每篇新闻做了正负语调词频的统计，从而构建正负语气指数，并以此情绪指数来研究 IPO 抑价率的变化。

不难看出，虽然文本数据所包含的信息在金融经济领域的实证研究中十分丰富，拥有极强的研究价值。但是现有的主要研究方法还是通过词频统计进行，这样的方法往往从一两个或者多个关键字词入手，而忽视了文章本身的全局信息，导致文本语义割裂，从而最终提取出的信息可能存在一定偏差，进而影响后续的实证研究。因此，在对金融文本进行分析时，我们需要引入一些自然语言处理领域目前较为先进的技术。

2.3 自然语言处理技术发展

自然语言处理（NLP）问题的提出，是为了使计算机能够理解人类使用的语言，即自然语言，并与计算机语言进行转换，从而帮助人

类解决一些文本数据方面的问题。其本质问题就是如何通过计算机对语言文字进行数学计算，亦或者是如何有效度量两个单词、短语、句子甚至是文档之间的距离。最早期用来处理自然语言处理问题的模型，往往从单个单词或者短语入手，通过计算编辑文字的经过的最优路径长度，来衡量来衡量不同单词之间的计算长度。但是这样的方法限制性太强，不够方便，因此后续的研究思路多是如何将单词转化为向量的技术，也就是 Word2Vec。

Salton（1971）首先提出了文本计数的方法，即用单词出现在每个文本中的次数来表达该单词的特征向量，而 Sparck Jones（1972）在此基础上考虑了泛用型单词的特征，将文本数量做逆差，引出了 tf-idf 的衡量指标，但其本质上依然是以词频统计为核心，割裂了文本本身的信息内容，并不能抽出大量的潜在信息。此后，20 世纪不断涌现出的针对自然语言处理问题的机器学习方法，包括常用的支持向量机，贝叶斯分类器，隐含马尔科夫链、条件随机场以及主题模型等，其本质都是在词频统计的基础上，将单词用不同的方法转换成可计算的形式。一直到 2013 年，Tomas Mikolov 提出了 word embeding（词嵌入）的方法，自然语言处理的算法与应用开始腾飞。词嵌入的方法将每个单词映射到一个密集的向量空间中，要求在现实中语义相近的单词，在向量空间中，其余弦相似度也必须足够大；现实中语义较远的单词，在向量空间中他们的位置也要足够远。在此基础上，科研人员们将向量化后的词向量输入到机器学习或者深度学习的模型中，并由此完成自动问答、机器翻译、自动摘要、语义提取等多种功能。

而从 2000 年以来，自然语言处理问题中的语义提取（情感分析），也逐渐变成了一项主流的应用场景。Pang 和 Lee(2004)对三百多篇自然语言处理方向的论文进行了广泛的调查，发现了情感分析问题的共同挑战，以及意见挖掘（语义挖掘）工作的的主要任务与流程，并将其总结为意见提取、情感分类、极性确定和总结。而 Tang 等人讨论了与意见挖掘相关的四个问题，即主观性分类、单词情感分类、

文档情感分类和意见提取。对于主观性分类，他们强调了一些方法，如相似性依赖、NB 分类器、多重 NB 分类器和基于切割的分类器。O'Leary 提出了一个关于博客挖掘的调查，其中包括对博客搜索和挖掘的介绍，要分析的博客类型，要从博客中提取的意见的单位和类型，以及它们的应用。Montoyo 等人列出了一些开放性问题以及迄今为止在主观性分析和情感分析领域取得的成就。Tsytsarau 和 Palpanas 通过关注意见挖掘、意见聚合(包括垃圾邮件检测和矛盾分析)，提出了一个关于 SA 的调查。他们对意见挖掘方法进行了比较，这些方法被应用于一些常见的数据集。Liu 介绍了 SA 和意见挖掘方面可能存在的不同任务和发表的作品。列出的主要任务有:主观性和情感分类、基于方面的 SA、情感词典的生成、意见总结、比较意见的分析、意见搜索和检索、垃圾意见检测和评论质量。Cambria 等人指出了 SA 中涉及到的与当前需求有关的复杂问题，以及未来可能的研究方向。他们还列举了一些开放性的问题，如组成声明的 SA、自动实体识别、同一评论中的多实体讨论、讽刺性检测和更细层次的主观性分类。最近，Medhat 等人提出了一个关于特征选择和情感分类方法的调查。他们对特征选择方法(主要是点相互信息和 Chi-square)进行了非常简要的描述，并对情感分类方法和相关论文进行了详细讨论。除了这些论文，该领域还有大量的工作被报道，研究界已经创建了许多词汇表，以评估新设计的情感分析算法。特别是在过去的四年里，研究人员主要关注的是微博社区，它已被成功地应用于市场预测、社会广告和票房预测。但是与传统金融经济学中的情感分析不同，运用 NLP 技术进行金融新闻情感分析，旨在从文本中识别市场信心指标，而这些指标通常来源于投资者的判断或市场运动。

部署最简单的情感分类方法依然是基于统计的机器学算法，常用的诸如朴素贝叶斯分类器（Naïve Bayes），最大熵方法和支持向量机(SVM)。当前文本情感倾向性分析主要有基于统计学方法和基于词向量嵌入两大类。(林政等；2012)将情感、位置和关键词等属性作为抽取关键句的因子，然后把关键句群进行有监督和半监督的情感

分类，该方法效果较好，但抽取关键句方法有待完善，因为这样的方法更加依赖于作者的主观意向，对单个词句的出现频率、位置进行统计并不能使模型有效的理解单词的语义。而(Fan X et al.；2012)利用含有否定词表、倾向性词表、程度词表的情感词表训练文本进行特征扩展，将单词转化为密集向量，在文中两个单词的意思约靠近，两者向量的余弦值应当越接近 1。该方法相比其它方法可以获得比较好的结果，但没有考虑上下文信息。但文本数据往往体量十分巨大，受限于本身的结构设计，传统的机器学习方案很难在大型的语言应用场景中学习得到很好的知识。不同于基于统计学理论的机器学习模型，深度学习算法虽然因为他的黑盒特性而被人所诟病，但是在 AI 生成领域，它拥有不可替代的优势。它通过学习数据的潜在层次或特征向量，从而拟合出最优的预测结果。其中最基本的模型莫过于卷积神经网络（CNN）与循环神经网络（RNN）。

Kim(2014)将每一个句子输入到卷积神经网络的模型当中，提取句子级别的信息，并完成文本分类任务，这其中也包括情感分类。Johnson 和 Zhang（2017)利用更加深入的词语级别 CNN 来捕获文本的全局特征。与之不同的是，Zhang et al（2015)提出了一个单词级 CNN，取得了不俗的结果。Tang et al（2015)通过利用用户和产品信息进行文档级别的情感分类。

但是文本本身存在一个阅读的先后顺序，这个特性导致他可以与时间序列数据进行对比。因此不断有自然语言处理领域的研究人员将目光投放到循环神经网络上来。但是当网络模型的输入域输出都是序列的时候，RNN 的第一个问题也就凸显了出来，即当输入输出序列不等长的时候，如何处理文本数据。这个问题经常出现在机器翻译的应用场景中，当我们要求模型将一句话从一种语言翻译成另外一种语言时，这两句话的长度大部分情况下是不相同的，这就导致模型无法自由的输出。Ilya Sutskever（2014）提出了 Sequence to Sequence 的模型，使用两个 RNN 的结构，保证不等长的序列也能被串接起来。同年，Kyunghyun Cho（2014）页提出了相似的概念，并

将其定义为编码器-解码器框架。在此基础上，注意力机制(Attention)首先被用于机器翻译(Bahdanau et al, 2015)，这使得RNN架构的模型在进行文本数据的学习时，能顾成功注意到文本中的重要信息。接着(Yang et al；2016)和(Ma et al；2017)在文档级上，将其改编应用于用户-产品评论情感分类。针对中文金融新闻，(Luo et al；2018)设计了一个分层模型，从多粒度中学习文档的表示，并提出了一种查询驱动(query-driven)的注意机制，以满足财务文档的独特特性。之后，(Anonymous authors；2020)使用(Malo et al；2014)创建的金融短语库和(Maia et al；2018)创建的FiQA任务情绪评分数据集，探讨了在金融情绪分类中使用和预训练、微调语言模型的有效性。

在各类算法不断变得复杂，各种附加的规则与专家知识争奇斗艳的情况下，谷歌（2017）发表了《Attention is All Your Need》，至此，Transformer模型框架横空出世，以其优秀的性能打败了自然语言处理领域的所有子任务。

但由于 Transformer 的网络结构复杂，参数量极大因此在训练时，需要大量的监督数据;特定领域中的自然语言处理和生成任务又需要代价高昂专业知识。而近些年逐渐成为主流的自监督学习（SSL）框架又恰好可以解决这个问题。Google 在 2018 年提出了 BERT 模型(Bidirectional Encoder Representation from Transformers)，通过英文文本预训练加微调的方式，在各式各样的下游任务中取得不错的成绩。在此基础上，(Y Cui et al.；2021)针对中文在语法和分词上区别于英文的独有特性，设计了 Whole-Word-Mask 预训练任务，得到了中文的大型金融语言模型Chinese BERT。

由此可见，得益于 Transformer 模型框架与自监督学习训练框架的提出，当我们想要提升媒体情绪抽取的精度时，效果最好的方法应该是使用基于 Transformer 结构的模型在金融媒体情绪分类的场景中进行学习，这样，在不依赖于大量大量专家标记数据的情况下，也能从新闻文本中提取到有价值信息，从而为后续的实证研究服务。

3.金融新闻情感分类模型

3.1 模型架构

本文使用 BERT 预训练模型通过 Attention 机制对文本进行双向训练表征。BERT 模型拼接了 Transformer 架构重的编码器(Encoder)部分，本身具有极强的学习能力；另一方面，BERT 又基于 SSL 的框架，通过数据本身的特征构造自带标记的训练样本，从而极大程度的降低了数据的标记成本与获取难度。这样，本文训练出一个效果较好的新闻情感分类器的可能性将大大提高。

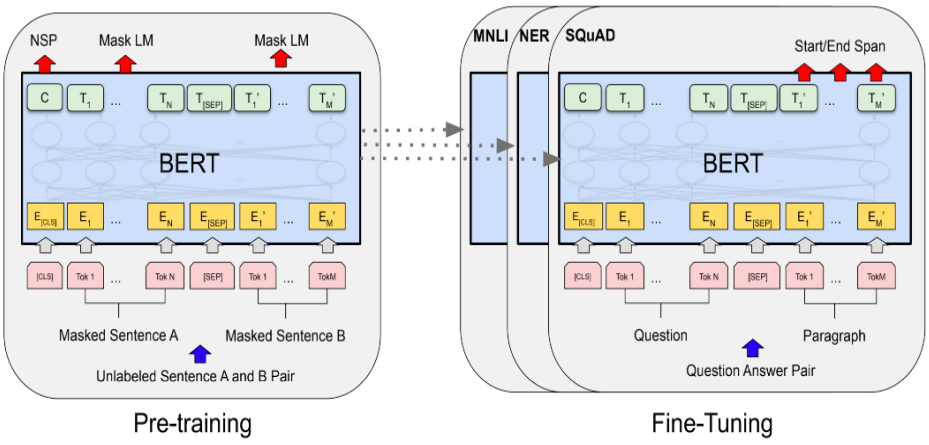

图 3-1　BERT 模型架构

*资料来源：HTTPS://GITHUB.COM/GOOGLE-RESEARCH/BERT

模型使用 Transformer 中的 Encoder 部分，并将 Encoder 层扩展为 6-12 个，每一层中保留 Transformer 的 Attention 机制。

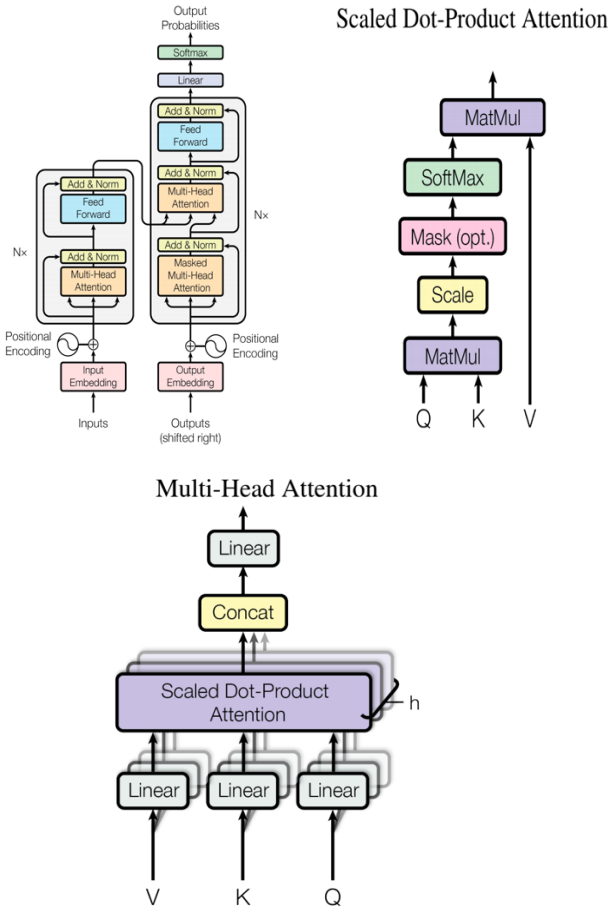

图 3-2　TRANSFORMER 模型架构

*资料来源：HTTPS://GITHUB.COM/GOOGLE-RESEARCH/BERT

我们假设向第一个 Encoder 中输入一个用 one-hot 编码（独热编码）表示的文本矩阵$S \in \mathbb{R}^{l\times m}$, l文本长度、m代表语料库单词总数。在 Encoder 层对自身添加注意力机制：

$$X = E \cdot S \tag{3-1}$$

$$Q = X^T \cdot W^Q \in \mathbb{R}^{l\times d_k} \tag{3-2}$$

$$K = X^T \cdot W^K \in \mathbb{R}^{l\times d_k} \tag{3-3}$$

$$V = X^T \cdot W^V \in \mathbb{R}^{l\times d_v} \tag{3-4}$$

$$Attention(X) = softmax\left(\frac{(X^TW^Q)\cdot(X^TW^K)^T}{\sqrt{d_k}}\right)V \in \mathbb{R}^{l\times d_v} \quad (3-5)$$

这样，我们能将一个巨大语料库中的文本单词，映射到一个密集的向量空间中，通常映射过后每一个单词将用一个 768 维或 1024 维的向量进行特征表示。

但是，原生的 BERT 模型在做中国金融市场的新闻情感分类时有两个明显的问题。一是 BERT 的分词方式是为英文设计的，它依据空格进行分词并将每一个英文单词进行编号，为后续特征提取做准备。在进行预训练时，BERT 会将分词后的每个单词以 15%的比例进行随机遮挡，并将遮挡后的句子输入模型预测出被掩码前的结果，以此进行模型训练。总所周知，英文单词通常以空格区分，每一个单词能表达的意思可能是多元的。但如果将每一个单独的汉语短语拆开进行是显然不适合中文的语法规则与习惯的，因为大部分时候，中文单个单词的意思与整个词语甚至短语要表达意思相去甚远，仅仅对一个词语中的每一个字做处理来进行预训练是显然不合理的。因此(Y Cui et al.；2021)依据中文的语言习惯与行文逻辑，使用全词掩码（Whole Word Mask）的模式，训练出了适用于中文场景的 BERT 预训练模型。

表 3-1 全词掩码规则

	中文句子
原始句子	我们使用全词掩码模型对中文金融新闻进行预训练。
分词结果	我们 使用 全 词 掩码 模型 对 中文 金融 新闻 进行 预训练 。
模型输入	我 们 使 用 全 词 掩 码 模 型 对 中 文 金 融 新 闻 进 行 预 训 练 。
原始掩码	我 [M] 使 用 全 词 掩 码 模 型 对 中 [M] 金 融 新 闻 进 行 预 训 练 。
全词掩码	[M] [M] 使 用 全 词 掩 码 模 型 对 [M] [M] 金 融 新 闻 进 行 预 训 练 。

*其中[M]代表被随机遮掩的单词。

另一方面，BERT 等开源的预训练语言模型都是在通用语料库上进行训练以及测试的，这样的模型在日常生活的使用中具有较强的效果，但是在金融这类专有名词较多的应用场景中，或许需要一个专有的金融 BERT，才能达到较好的效果。例如，在金融领域，不少新闻、研报已经论坛评论可能都会出现“杠杆”这个单词，在金融中往往代表撬动数倍资金的风险行为，但是一个通用的 BERT 模型可能会将其理解成物理上的工具“杠杆”，因为在训练的过程中，神经网络学习到的将“杠杆”理解为后者的次数更多。因此，本文将结合上述两点，使用(Y Cui et al.; 2021)提出的全词掩码机制，在没有标记的金融文本语料库上对 BERT 模型进行预训练。

3.2 金融 BERT 预训练

本文采用 CSMAR 上下载的新闻数据进行预训练，筛除过短（低于 128 字）与过长（高于 1024 字）以及重复内容的数据之后，共计余下 700 万条金融新闻数据。同时，文章选取 Whole-Word-Mask 的预训练方法与 HuggingFace 的大模型框架，在上述金融语料库中进行预训练。

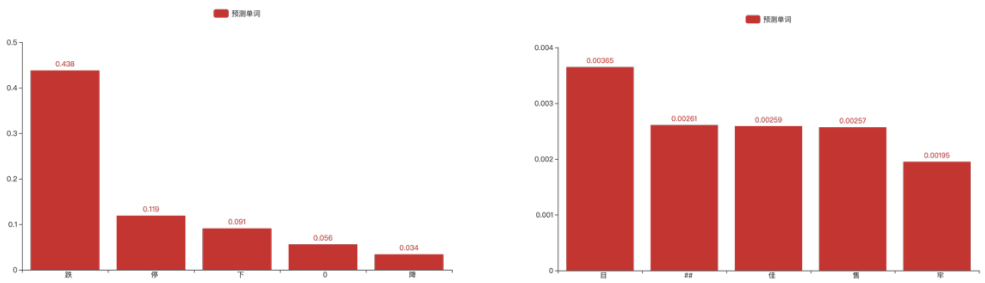

图 3-3　中文金融 BERT（左）与通用 BERT（右）预测对比

*上图展示了中文 FIN-BERT 模型与通用 BERT 模型在预测被遮挡单词出现概率，图中展示了其中概率最大的前五位，横轴表示预测出的字符，纵轴表示预测概率。

图 3-3 中左侧为本文训练出的金融 BERT 模型，右侧为(Y Cui et al.; 2021)开源的同样使用 Whole-Word-Mask 进行预训练的通用中

文 BERT 语言模型。当我们将金融中常见的一种句式“招商银行股价近期呈现下跌趋势”进行遮挡并让模型做出预测时，我们的模型能够以较大的优势预测出下跌，而通用的中文 BERT 却是将下与常见字进行组词，以大致相同的概率预测与“下”相关的搭配。在金融文本中训练的 BERT 模型，其结果将更适用于金融场景的应用，在此处可见一斑。

3.3 模型微调

为了使模型能够对金融文本的情感做出准确判断，本文还需要对模型进行微调。针对下游任务的微调，本文一共考虑了三种网络结构，一方面，文本数据本身存在先后顺序，在处理是可以当做时序数据进行操作，因此我们首先考虑了时序中的经典模型 LSTM 与 GRU，仅考虑一层隐藏层，且隐藏层神经元个数为 128；此外，鉴于预训练模型 BERT 其网络结构本身就已经足够庞大，因此再接入复杂网络可能会产生过拟合问题，因此第三种模型本文考虑了直接接入线性分类器。

在数据选取方面，我们选择从 ChinaScope 中抽取带有情感标记的金融新闻数据，抽取规则如下：

考虑到常见的金融新闻长度，我们将新闻总篇幅控制在 500 字左右。同时，统计了 ChinaScope 中所有新闻的字数信息，发现大部分新闻长度为 256-512 个字，因此，本文最终选取长度为 390-512 个字的新闻；

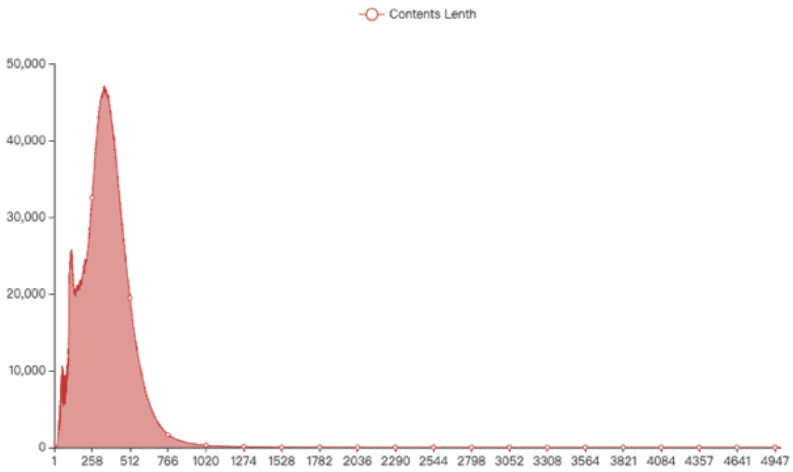

图 3-4　新闻长度统计

其次，在统计新闻来源之后。本文发现来自于一些自媒体或者公众平台的新闻较多。因此，本文筛去了“新浪网”、“网易网”等综合性媒体的报道，因为他们的新闻可能在金融领域质量不高，还夹杂着诸如评论之类的无效信息；而诸如“金融时报”、“格隆汇”之类的媒体，他们的新闻内容以股权变动、公司新闻、证监会公干等与金融高度相关的内容为主，因此更加受到我们的关注。

图 3-5　新闻来源统计

此外，ChinaScope 将金融新闻的情感极性分为 3 类，分别是中性、积极、消极，为避免样本不均匀而导致训练出现偏差，本文在三种新闻上均匀取样，最终每种感情的新闻选取 5000 条，构成最终的 15000 条训练数据。

本文将 70%的数据划分为训练数据集，10%的数据划分为验证数据集，余下 20%作为测试数据集。此外，由于 LSTM 与 GRU 模型要求前后输入的序列长度一致，因此，我们将对输入的新闻做统一的截断或者补余。本文共选取了三种截断长度，分别为 32,128,512，其中 32 为大部分金融新闻的摘要长度，128 通常为新闻的首段长度，512 几乎囊括所有新闻的全文。另一方面，由于选取的截断长度的不同，带来的模型训练与特征数据存储成本也是差异巨大的，这一点，我们将在本章第四节结果分析部分进行说明。此外，经过小规模实验的多次调试，我们将 learning rate（学习率）定为 0.00001，模型训练的 Batch Size（单次输入数据的批次大小）设置为 30。

3.4 训练结果及说明

我们分别使用本文训练的中文金融 BERT(下述 Model1)、哈工大开源的中文 whole word mask BERT(下述 Model2)、百度开源的中文通用 BERT(下述 Model3)，以及熵简科技开源的基础版中文金融 BERT(下述 Model4)，在上述小节的条件下进行模型的微调，结果表 3-2 所示。

表 3-2　模型微调结果对比

Finetune-Model	Padding		
	32	128	512
Model1-Linear	0.885(0.725)	0.873(**0.803**)	0.89(**0.819**)
Model1-LSTM	0.86(0.689)	0.893(0.757)	0.901(0.769)
Model1-GRU	0.876(0.688)	0.905(0.754)	0.903(0.759)
Model2-Linear	0.868(0.679)	0.882(0.781)	0.885(0.807)

Finetune-Model	Padding		
	32	128	512
Model2-LSTM	0.855(0.687)	0.852(0.717)	0.909(0.754)
Model2-GRU	0.873(0.683)	0.87(0.715)	0.915(0.772)
Model3-Linear	0.769(0.567)	0.792(0.665)	0.83(0.696)
Model3-LSTM	0.722(0.63)	0.771(0.66)	0.802(0.675)
Model3-GRU	0.764(0.571)	0.809(0.634)	0.826(0.66)
Model4-Linear	0.816(0.637)	0.879(0.685)	0.881(0.70)
Model4-LSTM	0.783(0.645)	0.801(0.673)	0.813(0.681)
Model4-GRU	0.779(0.593)	0.81(0.645)	0.816(0.675)

*上表展示了模型微调后的情感极性分类准确率，括号外为训练集的准确率，括号外为测试集的准确率，两者差值可衡量过拟合程度。模型精确度评价指标选取测试集准确率。

表 3-2 展示了 12 组微调模型组合在三种截断（或补全）规则下的模型分类效果，其中括号外侧为训练集的平均分类准确率，括号内侧为测试集的最终实验分类准确率。首先对比同一预训练模型下的不同微调模型，不难看出四组预训练模型下，接入线性分类器的效果都是最好的，由训练集与测试集的准确率对比可以得知，接入 LSTM 与 GRU 模型后，由于网络结构较为复杂，产生了过拟合的情况，导致训练时效果极好，准确率基本维持在 90%，但测试集的准确率相较于线性分类器却大福下降。

其次，再纵向对比四组预训练模型，不难看出本文训练出的中文金融 BERT 模型（Model1）在各组下游模型的对比结果中几乎全部取得优势，而使用相同训练方式的 Model2 却因为预训练语料不是金融新闻文本，而始终在模型精度上维持着一定的差距。Model4 也与前两个模型使用了相同的预训练方法，且也是在金融文本上进行训练，但因为其仅使用了 6 层的 Transformer 编码层，最终生成的特征向量为 768 维，而 Model1，Model2 沿用了最大版本的 BERT，使用了 12 层 Transformer 编码层，最终特征向量为 1024 维，保留了尽量多的文本信息，因此在性能上大幅优于 Model4。而百度开源的模型

Model3，也同样仅释放了基础版，特征向量维度为 768 维，且未能在金融语料上进行预训练，因此可以看出其实验效果明显弱于前两组模型。

最后从新闻文本截断或补全的长度来看，当序列长度选取为 512 时，模型几乎能够得到整个文本的信息，因此，不难理解在所有模型下，序列长度为 512 时分类效果最好。不过值得注意的是，当选取序列长度为 128 时，模型分类效果虽然略有下降，但依然维持在一个较高水平，且相比序列长度为 512 时的情景，下降并不严重。这是因为在进行研究时，我们筛选的新闻样本数据几乎都为官方权威媒体的撰稿，他们一般行文十分严谨，有固定的写作范式，这使得一篇中短篇新闻的前 128 个字几乎能够囊括整篇新闻的主要信息，纵使有少量的信息缺失，也不会对模型精度造成致命影响。然而，当序列长度下降到 32 时，模型精度却有了明显的下降，但是 32 个字也应该刚好足以囊括新闻一两句话的摘要，捕捉到了主要信息，模型精度不应当有如此大的差别。于是我们再重新观察样本数据，发现在这类官方媒体撰写的新闻段首，多有不同形式的单位或者作者信息，如“本报讯”、“×××网报道”，这类信息不仅挤占了主要信息的输入空间，还会对本就不长的文本序列造成干扰，导致效果大幅降低。

	emotionIndicator	newsSummary
0	0.0	百事公司任命谢长安为大中华区首席执行官11月1日，百事公司宣布任命谢长安女士为大中华区首席执...
1	0.0	格隆汇6月16日丨*ST华塑(000509.SZ)公布，截至本预案公告日，公司股本总额为8....
2	0.0	2021年11月17日沃尔核材(002130,股吧)（002130）发布公告称：万和证券刘圣...
3	0.0	在官网中，英伟达将Omniverse描述为“一个将3D世界连接到共享虚拟世界的平台”，可完成...
4	0.0	中国网地产讯 2月3日，中国银行发布公告称，自2月2日起至国家确定的疫情结束之日止，中国银行...
...	...	...
9995	2.0	4月17日早间消息，据外媒Macrumors报道，分析师杰夫·普在最新的报告中表示，由于搭载...
9996	2.0	新浪财经讯 12月13日消息，交行晚间公告称，选举任德奇为交通银行股份有限公司董事长。任德奇...
9997	2.0	一、减持计划的主要内容航锦科技股份有限公司（以下简称“公司”）于2021年6月19日披露了《...
9998	2.0	4月24日，在黑龙江省政府新闻办举行的绥芬河口岸跨境输入疫情防控情况第五场新闻发布会上，牡丹...
9999	2.0	共达电声终止重组,因万魔声学今年业绩不确定性大且国内证券政策已变 来源：爱集微集微网消息，7...

图 3-6　模型微调样例数据

上述实验结果基本验证了本文之前提出的第一个研究意义，即通过将 Transformer 与 Self-Supervise Learning 结合的 BERT 模

型，可以在大幅减轻学术研究和量化分析的数据标注成本的基础上，依然取得一个很好的模型效果。我们可以选取尽量多层的通用版本 BERT 模型，使用低成本的无监督金融文本数据进行预训练。另一方面值得注意的是，在实际进行学术研究或者量化分析时，我们并不是一定要选取最大的序列长度以保证模型吸收尽量多的信息。因为虽然输入的序列长度越长，模型得到的信息越充分，最终的模型精度也就越高，但是这无疑将几何倍的消耗我们的研究成本。因此，我们需要在模型精度以及算力、数据成本上做平衡，确保文本的主要内容输入能够进入模型即可，这样不仅能够保证模型精度，还能极大的缩小成本。在训练时我们注意到，同样使用线性分类器的情况下，当序列长度为 128 时，模型遍历整个数据集花费不到 40 秒，而当序列长度提升至 512 时，模型遍历整个数据集的时间超过 8 分钟，扩大了超十倍。如果想要提前生成模型的特征向量便于后期随取随用，序列长度为 512 的文本特征也会为数据的存储和调用带来不小的压力。以 1024 维的特征向量为例，当序列长度为 512 时，1 万篇新闻生成的特征矩阵大小超过 20 个 GB，如果研究或分析的新闻数量超过百万级，这无疑是一个巨大的负担，反而增加了在金融市场挖掘金融文本信息的难度。综上所述，后续研究中，在保证模型精度的基础上，本文选择了 Fin-BERT-Linear 的模型，在序列长度为 128 的基础上，完成了后续的媒体情绪因子生成与构造。

4.媒体情绪因子的定价分析

4.1 模型选择与数据选取

4.1.1 模型与假设

本章节中，我们将使用因子模型论证情绪因子的定价能力，文章将在三因子模型的基础上，构造新的金融媒体情绪因子，看其是否能解释截面的收益率预期差异。本文首先考虑了最为经典的 Fama-French 三因子模型（Fama et al.，1993），但不同于美股市场，Fama 的三因子模型在 A 股市场上多次吃瘪。Liu et al.(2018)研究发现，这可能是由于中国市场特有的 IPO 监管政策，导致一些公司借壳上市，对证券市场造成了壳价值的污染，因此研究时必须抛弃市值最小的 30%的公司。本文将 Liu et al.(2018)的中国版三因子模型作为基础模型，再加入自身的媒体情绪因子生成四因子模型。本文认为，由于金融媒体报道时本身具有一定的观点输出行为以及相应的立场，甚至影响投资者对市场的看法，导致资产在一定程度内偏离了正常的水平。由此，本文提出假设：

H1a：金融媒体对上市公司的倾向性报道将会影响其未来股价，且报道的情感越正面，公司股票收益率越高；

但是从风险补偿的角度来看，当一只股票的媒体情绪为负时，其在市场上将会释放出消极信号，导致持有该资产的投资者风险上升，那么持有该资产的投资者将会获得额外的超额收益作为补偿。因此我们提出假设：

H1b：金融媒体对上市公司的倾向性报道将会影响其未来股价，且报道的情感越负面，公司股票收益率越高

另一方面，本文引入金融媒体情绪因子，若其能够显著影响投资组合的超额收益率，则说明该因子拥有三因子模型没有的定价信息，

相应的四因子模型将会有更低的定价误差α以及对应更低的 t 检验值。由此，我们提出假设 2

H2：金融媒体情绪因子能够增强三因子模型对股票收益率的预测能力。

4.1.2 数据选取

本文选取了 CSMAR 上的全部新闻数据以及部分 ChinaScope 上的新闻数据构造媒体情绪指标，筛选出能匹配到对应股票和公司的新闻共 181 万条，但由于 2014 年之前的新闻数据存在较多缺失，许多月份找不到新闻，因此我们将时间窗口缩短至 2014-01-01 至 2022-01-31，共计 1,732,665 条新闻数据。

另一方面，依据 Liu et al.(2018)的研究，本文剔除了 2014-2021 年间总市值位于后 30%的股票，并剔除剩余股票中仍留存的 ST 股票。此外，本文还在 CSMAR 上下载了 2014.01-2022.02 时间段内共 97 个月的月度市值数据、月度市盈率数据值，以及对应的下一个月的个股月收益率、月度无风险收益率数据

4.2 媒体情绪指标

4.2.1 媒体情绪指标的构造

首先使用 Fin-BERT-Linear 模型判断筛选出的新闻的情感极性，最终结果发现约 44.5%的新闻被判定为积极情绪，35.7%的新闻被判定为消极情绪，其余 19.8%的新闻被判定为中性情绪。

表 4-1　新闻情感分类数量

积极	消极	中性
771036	618561	343068

文章借鉴 Garcia(2013)对新闻情绪的做法，将每月每只股票的对应新闻情感求均值（其中积极情感表示为 1，消极情感表示为-1，中性情感表示为 0）。但是发现求得的个股月度情绪指数出现大量为 0 的值。

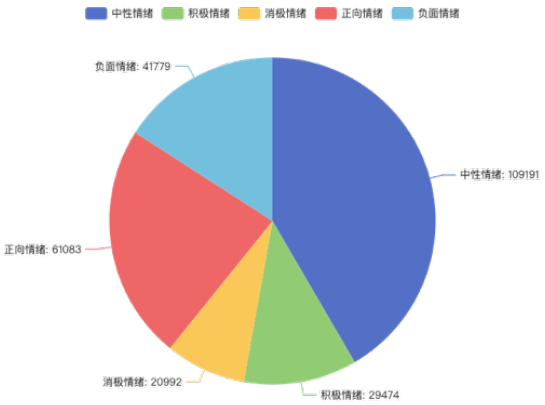

图 4-1　情感极性月度均值

其中，积极情绪和消极情绪分别指代情绪指标为 1 和-1 的情绪值，正向情绪指不小于 0 的情感极性，相应的负向情绪指情绪不大于 0 的情感。值得注意的是，本身只占 20%不到的中性情感在情感极性的月度均值中占比高达 45 以上，这是由于同一月中，单个股票可能同时出现相同数量但是情感极性相反的文章，导致求和结果为 0，使用该种方法构造的指数很难反应市场的情绪变化。

因此本文考虑了模型的最后一层，接入一个 Softmax 归一化层，不直接输出最终的情感极性，而是输出各个预测为各种情感的概率，最终的情绪指数为：

$$senti = p(senti = \text{“}pos\text{”}) \times 1 + p(senti = \text{“}neg\text{”}) \times (-1) + p(senti = \text{″neu″}) \times 0 \quad (4\text{-}1)$$

根据公式不难看出，该情感指数越大，新闻所表述的情感就更加积极；情感指数越小，新闻所表述的情感就更加消极。另外，当某一个月中某只股票没有新闻报道时，本文将在当月剔除该股票，以免在构造投资组合创建因子时，额外导入过多中性情感。

4.3 引入情绪因子的四因子模型

4.3.1 因子构造

本文因子的构造方法依然沿用了 Fama&French（2015）推荐的方法，使用2×3的分割方式来构造因子，得到两个2×3的投资组合，即 Size-EP 与 Size-Senti。其中将股票市值按大小排序，前 50%标记为大规模（B），后 50%标记为小规模（S）；将市盈率低于 30%分位点的标记为低（G），高于 70%分位点的标记为高（V），中间部分标记为中（M）；针对情感因子，因其本身有三种分类，因此我们也按照 EP 因子的方式分为三组，高于 70%分位数的标记为积极（P），低于 30%分位数的标记为消极（N），其余标记为中性（U）。上述三个因子的具体计算方式为：

$$SMB_{EP} = \frac{SV + SM + SG}{3} - \frac{BV + BM + BG}{3} \quad (4-2)$$

$$SMB_{Senti} = \frac{SP + SU + SN}{3} - \frac{BP + BU + BN}{3} \quad (4-3)$$

$$SMB = \frac{SMB_{EP} + SMB_{Senti}}{2} \quad (4-4)$$

$$VMG = \frac{SV + BV}{2} - \frac{SL + BL}{2} \quad (4-5)$$

$$PMN = \frac{SP + BP}{2} - \frac{SN + BN}{2} \quad (4-6)$$

由公式（4-2）到公式（4-6）的方法，本文构造出了四个因子收益率的时间序列数据，观察其统计特征并通过其中的 t 检验统计量来判断该因子是否显著为 0。

表 4-2　因子描述性统计

	MKT	SMB	VMG	PMN
Mean	0.0622	0.0414	0.0183	0.0346

	MKT	SMB	VMG	PMN
Std.Dev.	0.453	0.114	0.046	0.111
t-stat.	1.21	3.22	3.51	2.76

通过对因子的描述性统计可知，四个因子中，MKT 因子的标准差最大，并且其在时序上的分布特征并不能看出显著不为零，这可能是因为 14 年以来 A 股市场上本身市场波动较大，从而导致该因子也极为不稳定。但是其余三个因子与市场因子相比之下，数值都较为稳定，且在 14-21 年末的时间序列分布上显著不为零。

表 4-3　因子间相关系数

	MKT	SMB	VMG	PMN
MKT	1	0.259	-0.33	-0.348
SMB	0.259	1	-0.45	-0.31
VMG	-0.33	-0.45	1	0.27
PMN	-0.348	-0.31	0.27	1

此外，通过计算四个因子之间的斯皮尔曼相关系数，可以看出各个因子之间并没有较强的相关性。其中最强相关性的一组因子为 SMB 与 VMG，这一结果也与 Liu et al.(2018)观察到的结果基本吻合。另外，从相关系数可以看出，市场因子与市值因子为正相关关系，但是两者相关系来看，并不存在较强的相关性，市盈率因子与金融媒体情绪因子之间相关系数也为正，但相关性只有 0.27，并不强烈。此外，市场因子、市值因子与市盈率因子、金融媒体情绪因子之间两两相关性为负向，说明当市场总体收益率越高时，市值较大的公司个股的投资组合往往收益率较大，而市盈率较大的公司个股组成的投资组合往往收益率较低，相关媒体情绪越积极的上市公司，其相关的投资组合收益率也较低。而当市场总体收益率较低时，市值较小的公司个股的投资组合往往收益率较大，而市盈率较小的公司个股组成的

投资组合往往收益率较低，相关媒体情绪越积极的上市公司，其相关的资产组合在这种情况下可能有较大超额收益。

但总体而言，四个因子的相关性系数并不高，也就是说四个因子作为回归分析的自变量时，变量之间不存在较强的多重共线性。因此，为了更有效的说明引入媒体情绪因子（PMN）后的四因子模型是否有显著的定价能力，以及 PMN 因子是否增强了对 Liu et al.(2018)的中国版三因子模型定价误差的解释力度，本文将在下一小节进行回归分析。

4.3.2 因子回归

本节采用线性回归的方式，主要对 Liu et al.(2018)的中国版三因子模型，以及本文引入媒体情绪因子后的四因子模型进行对比，验证模型在 A 股市场上的定价能力。

首先，我们对市场因子进行单因子的回归。

$$R_{it} = \alpha_{it} + \beta_i^{MKT} * MKT_t + \varepsilon_{it} \qquad (1)$$

随后，将其余三个因子与市场因子，分别组合，生成模型 2-模型 4。

$$R_{it} = \alpha_{it} + \beta_i^{MKT} * MKT_t + \beta_i^{\lambda} * \lambda_t + \varepsilon_{it} \qquad (2\text{-}4)$$

其中，λ_t代表规模因子 SMB、市盈率因子 VMG 以及金融媒体情绪因子 PMN 中的任意一个；而β_i^{λ}代表代表针相应的λ因子（即规模因子、市盈率因子以及金融媒体情绪因子）对个股或者资产组合i的因子负载能力。当回归后β_i^{λ}的在 t 检验上显著时，可以说明该因子λ对个股或者资产组合的超额收益有着十分显著的影响。最后，本文构造了基于市盈率因子的中国版三因子模型：

$$R_{it} = \alpha_{it} + \beta_i^{MKT} * MKT_t + \beta_i^{SMB} * SMB_t + \beta_i^{VMG} * VMG_t + \varepsilon_{it} \qquad (5)$$

其中α_{it}代表该因子模型在 t 时刻对个股或者资产组合 i 的定价误差，当回归后其在时间序列分布以及截面上的总体定价误差α显著为零时，说明该模型对相应市场上的个股或者投资组合的超额收益

有着较好的解释能力，模型中选取的因子能够解释市场上投资组合产生超额收益的绝大部分原因。而当α显著不为零时，则说明该模型仍然不能完美解释当前市场下个股或者投资组合产生超额收益的所有因素，依然存在其他的因子拥有额外的信息可以解释一部分收益产生的原因以及市场异象，此时查看单个因子是否有用可以回到上一部分阐述的方法中，观察相关因子回归系数的显著性，从而判断单个因子对解释市场上个股或者投资组合出现超额收益的原因是否有用。

最终，本文引入金融媒体情绪因子，构造 A 股市场的四因子模型：

$$R_{it} = \alpha_{it} + \beta_i^{MKT} * MKT_t + \beta_i^{SMB} * SMB_t + \beta_i^{VMG} * VMG_t + \beta_i^{PMN} * PMN_t + \varepsilon_{it} \quad (6)$$

在对比两个因子模型的优劣时，在线性模型中，常用的方法中一般会先度量回归直线对观测数据的拟合程度，也即可决系数（R^2），一般而言，当模型的可决系数越大时，说明该模型的线性拟合效果越好，模型的可解释性越强。但是可决系数这一指标实际上并不充分，尤其是在模型采用的因子数量不一致时，自变量越多的模型，其可决系数R^2自然也就越大，因此，我们还将需要一些更加具有说服力的结果。如前所述，通常验证因子模型是否能够解释当前市场上投资组合的超额收益时，我们往往观察模型总体定价误差α是否显著为零，如果不同模型的α值以及其显著性大小有明显的差异，则可以通过该指标进行比较

上述模型 1-6 中，下标t指代个股的以及因子收益率的时间节点，下标i指代个股的编号。不难看出，本文用于回归的个股数据既有时间序列数据，也包含了截面数据，因此在做回归分析时，既需要进行时序层面的线性回归，也需要进行截面数据的线性回归。此次研究中，本文采用了 Fama 和 Macbeth（1973）所使用的的回归方法，第一步本文通过时间序列回归的方式，将每一只股票i在时序上做一次

回归分析，求得个股收益率在该因子上的因子暴露β_i^{λ}，其中λ指代相应的因子。根据 Fama-Macbeth 的回归方法，我们还将在每个 t 时刻对该时刻下的个股数据做一次独立的横截面回归，并将总共的 T 次截面回归所获取的参数求平均值作为最终回归分析的参数值：

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\hat{\lambda}_t \qquad (4-7)$$

$$\widehat{\alpha_i} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\alpha_i} \qquad (4-8)$$

这种回归方法的巧妙之处在于将 T 期的回归结果看做为 T 个独立样本，从而他可以排除残差在截面上的相关性对标准误的影响。

通过以上方法对模型 1-6 进行回归结果如下

表 4-4　回归分析结果

Model	(1)	(2)	(3)	(4)	(5)	(6)
α	-0.124***	-0.056***	-0.082***	-0.075***	-0.014**	-0.00673*
	(-61.5)	(-16.7)	(-16.6)	(-15.8)	(-2.65)	(-1.67)
β^{MKT}	1.10***	1.01***	1.03***	0.987***	0.981***	0.975***
	(78.16)	(75.85)	(68.03)	(73.45)	(66.66)	(65.31)
β^{SMB}		0.643***			0.567***	0.499***
		(25.12)			(21.47)	(19.06)
β^{VMG}			-0.448*		-0.378*	-0.369
			(-1.72)		(-1.66)	(-1.43)
β^{PMN}				-0.482**		-0.385*
				(-2.07)		(-1.92)
N	180866	180866	180866	180866	180866	180866
R^2	0.112	0.174	0.162	0.154	0.341	0.387

注:*、**、***分别代表在 10%、5%、1%水平下显著相关

最终的回归结果如表 4-2 所示，首先，从单个因子的回归情况来看模型 1-4 的定价误差α均显著不为零，但所有因子都与股票的超额收益显著相关，其中市场因子与规模因子对资产组合的收益率有正向的影响，而市盈率因子与金融媒体情绪因子对资产组合的收益率均有负向的影响。但无论是哪个因子，最终的定价误差显著性都特

别高，这说明单个因子的信息量完全不足以解释资产产生超额收益的原因。

与单因子模型相对比的，我们的三因子模型得到了一定的提升，不仅拟合优度显著升高，而且定价误差α也大幅降低，更加接近于 0，与此相关的，定价误差α显著性也大幅度降低了，这说明我们的三因子模型中所包含的信息已经远远超过模型 1-4，能够不错的解释市场下投资组合产生超额收益的一部分原因了，但是仍然存在一些影响超额收益的因素是我们的三因子模型所无法解释的。

通过分析模型 6，即引入金融媒体情绪因子的四因子模型，可以看出除了市盈率因子以外，模型的各个自变量都是显著的，其中我们引入的金融媒体情绪因子无论是在单因子回归还是多因子回归中都有着极强的显著性，并且系数为负。这就能验证我们在之前提出的假设 H1b，即金融媒体的情绪能够影响资本市场上个股或者资产组合的预期收益，且当前情感越消极的股票组合或者资产组合，在未来越有可能获得高额的回报。对于这个结果，本文尝试从资产定价的角度进行分析，当金融市场上的媒体对一个公司发布了大量新闻的时候，这说明该公司的信息透明度是比较高的，如果对于该公司相关新闻的媒体情绪为正，则该种情绪也将影响到市场上的投资者，投资者对该公司的未来有一个较好的预期，从而使得公司的股票价格被高估，导致下一期的个股或者相关投资组合的预期收益率降低；而当与该公司相关的新闻的媒体情绪为负时，该种情绪影响到市场正常行为后，公司股票价格将被低估，导致下一期的个股或者相关投资组的预期收益率反而升高。并且，从风险补偿的角度来讲，当投资人持有的股票被市场上的媒体情绪影响而导致不被看好时，持有人往往需要承担更大风险，相应的，其风险补偿的预期收益率也就越高。

另一方面，通过对比三因子模型与四因子模型，我们可以发现引入金融媒体情绪因子后的四因子模型其拟合优度有一定程度的上升，但是从线性回归的统计学原理来看，当自变量因素增加后，其线性模型的拟合优度必然增加，可决系数R^2必定上升，因此从可决系数

R^2得出四因子模型由于中国版三因子模型这一结论并不充分。第二，通过观察三因子模型与引入微博情绪因子的四因子模型的模型总体定价误差α，可以看出，在数值本身方面，引入金融媒体情绪因子后的四因子模型，其模型总体定价误差相比三因子模型的定价误差大幅下降，更加接近于 0。其次，在定价误差α的显著性上，四因子模型相比三因子模型也有一定程度的下降，三因子模型在 5%的水平下显著相关，而本文的四因子模型在 10%的显著水平下相关，这说明四因子模型的定价误差维持原假设$\alpha = 0$的概率更大，这个现象可以在一定程度上说明，加入金融媒体情绪因子的四因子模型，与中国版三因子模型相比，对 A 股市场上个股与投资组合的超额收益有着更好的解释力度，可能包含了更多的信息来解释收益产生的原因。

但这依旧不足以支撑我们的假设 H2。为了探究假设 H2 是否成立，即引入新的因子是否增加的新的定价信息，本文使用均值方差-张成检验，查看金融媒体情绪因子中，是否包含三因子模型中市场因子、市值因子以及市盈率因子所不包含的信息。

Huberman and Kandel(1987)使用均值方差-张成检验来考察用 n 个资产构建的均值方差有效前沿能否包含某个新资产。因此，本文考虑了金融媒体情绪因子与市场因子 MKT、市值因子 SMB、市盈率因子 VMG 的回归模型如下：

$$r_t^{PMN} = \alpha + \beta_1 * r_t^{MKT} + \beta_2 * r_t^{SMB} + \beta_3 * r_t^{VMG} + \epsilon_t \qquad (7)$$

若α显著不为零，说明金融媒体情绪因子中含有三因子所不包含的定价信息，则我们引入这个因子是有意义的。回归结果如下：

表 4-5　均值方差张成检验结果

	α	*MKT*	*SMB*	*VMG*
para	*0.017**	*-0.097****	*-0.863***	*-0.139***
t	*2.25*	*-9.75*	*-2.10*	*-2.32*

注:*、**、***分别代表在 10%、5%、1%水平下显著相关

由表 4-5 不难看出，回归截距项在 5%的水平下显著不为零，这说明我们构造的金融媒体情绪因子确实存在额外定价信息，是其余三个因子解释不了的，那么，将该情绪因子引入模型确实提供了增量信息，是有必要的。最终，我们的假设 H2 也得到了验证。

4.3.3 分组测试

在本小节中，本文依旧沿用 Fama 的方法，分别使用市值因子（SMB）-市盈率因子（VMG）与市值因子（SMB）-金融媒体情绪因子（PMN）构造 5×5 投资组合，观察在不同分层情况下投资组合的收益率大小情况，并通过定价误差的显著性，检验四因子模型在不同投资组合下的定价能力。投资组合的划分如下表示：

表 4-6　不同分组的投资组合收益率

规模-市盈率	高	2	3	4	低
大	-0.019852	-0.007012	0.012817	0.025673	0.030822
2	-0.017482	-0.004187	-0.004653	-0.00809	-0.006792
3	0.001742	0.016308	-0.00876	-0.00278	0.005324
4	-0.014214	0.000467	-0.008311	-0.004516	-0.007749
小	-0.034618	-0.01192	0.003762	-0.006816	0.177406
规模-媒体情绪	积极	2	3	4	消极
大	0.013768	0.017935	0.015634	0.026391	0.0279623
2	0.012674	0.016006	0.017107	0.023703	0.0259865
3	0.0083219	0.0116473	-0.019432	-0.001892	0.0163283
4	-0.007671	0.009278	-0.171341	0.002891	0.0154234
小	0.003518	-0.047568	-0.001284	0.007898	0.0163471

在规模-市盈率分组中，当投资组合的规模较大时，市盈率上升，组合的收益将降低，而当规模降低时，这条规律就不再适用，此时随着市盈率的上升，投资组合的收益变化并不明显，但是当规模一定时，低市盈率的投资组合收益一定高于高市盈率的投资组合收益。这说明市场在规模-市盈率的分组中，规模效应并不明显，而当投资组

合所包含的资产规模都较大时，市盈率变化引起的预期收益变化较为明显，即存在市盈率效应，但这种效应只存在于规模较大的投资组合中，当投资组合包含的资产规模较小时，其市盈率效应就不明显了。

在规模-媒体情绪分组中，同样的，当投资组合的规模较大时，投资组合的媒体情绪越消极，其预期的超额收益率越高；当投资组合的规模较小时，随着投资情绪越来越消极，收益率并不线性上升，但是最消极的投资组合收益率一定高于同等规模水平下的最积极投资组合收益率。这或许能从两个方面来进行解释，其一，因为处于中性的新闻数量在中间中间的投资组合中占比较大，媒体情绪的极性都不强，使得分组时未能有效区分开；其二，当投资组合中的资产对应公司规模越大时，其媒体关注度往往越高，那么媒体对其进行相应报道的新闻数量也就越多，在进行投资组合的分类时，这种规模较大的资产在情绪指标的分类上往往更容易区分。另一方面，在规模-媒体情绪分组中，同样媒体情绪水平下，规模越大投资组合，其预期收率越高；规模越小投资组合，其预期收率越低，这说明 A 股市场在规模-金融媒体情绪的分类上，存在比较明显的规模效应。接下来，我们对分组后的每一个投资组合使用四因子模型进行时间序列的回归分析，探究我们的四因子模型在各个投资组合中的定价能力。

表 4-7　分组回归结果分析

	系数					t 值				
	大	2	3	4	小	大	2	3	4	小
	α					$t(\alpha)$				
高	-0.11	-0.12	-0.09	-0.11	-0.13	1.34	1.23	-1.04	1.33	2.01
2	0.06	-0.11	-0.13	-0.07	0.06	0.65	-1.07	1.32	-2.21	-1.58
3	-0.12	-0.10	-0.09	-0.07	0.09	-1.12	0.94	1.16	-2.57	-0.62
4	0.08	0.12	0.43	-0.17	0.11	-2.56	1.15	2.11	-0.98	-1.29
低	0.16	-0.04	-0.3	0.07	-0.26	-0.82	1.81	-2.23	-1.71	1.54
	α					$t(\alpha)$				
积极	0.21	-0.13	-0.16	-0.17	0.09	-1.46	-1.52	1.17	-0.75	1.41
2	-0.19	-0.03	-0.14	0.06	0.21	2.21	-3.83	0.45	0.72	-1.97
3	0.18	0.19	-0.12	0.05	0.17	2.36	-0.57	0.55	1.38	-1.23
4	-0.25	0.16	-0.26	-0.07	-0.11	-1.87	-1.21	1.49	-1.66	-1.73

消极	0.11	-0.15	-0.19	-0.08	-0.23	0.64	-1.85	1.33	-1.67	-1.84

可以看出，按照规模-市盈率分组的投资组合，在大部分情况下，其定价误差依然显著不为零；但是使用规模-媒体情绪分组的投资组合，却有不少实验在 10%及以上的情况下，其定价误差不再显著不为零。这与我们在上一节中的回归可以看出类似的结论，即我们的金融媒体情绪因子在这类场景下的定价能力可能优于市盈率因子。

进一步观察可以看出，效果较好的投资组合基本集中在规模较小，以及媒体情绪较为消极的投资组合中。这可能是因为，第一本身消极的媒体情绪更容易引起投资者的关注，从而对市场产生更大的影响。正向的新闻媒体情绪虽然也为影响投资者对未来的预期从而影响投资者决策甚至公司管理运营的决策，但是其影响较为轻微，投资者很少因为某一篇或者某几篇较为正面的报道就对大幅度改变自身的投资策略，大量购入相应的投资组合，企业经营者也不会因为相同的原因就改变公司策略以此博取媒体的好感。但是相反，负面的金融市场媒体情绪，可能会使得投资者与上市公司更为紧张。股票投资者可能因为投资该公司的资产风险较大而更改投资策略，上市公司为了尽快的消除媒体情绪带来的负面影响，也可能会在短时间内快速改变经营或者营销策略，造成更大的股价波动。第二，对于规模较大公司，市场对其的信心可能本就相比更大一些，当金融市场上出现不太强烈的媒体情绪难时，这种情绪以影响其在资本市场的表现。另外，这类公司往往在市场的信息披露也相对比较充分，信息透明程度也相应较高，当这类公司在市场上出现负面的媒体情绪以后，投资者往往还有别的信息源头可以对这类公司进行分析，也就是说，大规模的公司收到市场上媒体情绪影响的程度可能更小。

4.4 稳健性检验

4.4.1 媒体情绪因子在不同市场下的定价能力

为了验证本文所构建的四因子模型的稳健性，我们还在不同的市场上对模型进行了 Fama-Macbeth 回归。我们将市场分为沪市和深市，其中沪市共有相关新闻数据 655998 条，其余新闻全部为深市股票相关的新闻，可见深市的媒体情绪总体强于沪市。回归结果如下：

表 4-8　不同市场回归结果

	模型（6）沪市	模型（6）深市
α	-0.0076**	-0.0036*
	(-2.28)	(-1.74)
β^{MKT}	0.913***	0.899***
	(67.53)	(67.62)
β^{SMB}	0.672**	0.578***
	(22.24)	(18.97)
β^{VMG}	-0.469**	-0.38*
	(-2.17)	(-1.77)
β^{PMN}	-0.475*	-0.572**
	(-1.88)	(-2.13)
R^2	0.379	0.415

注:*、**、***分别代表在 10%、5%、1%水平下显著相关

可以看出，首先，深市和沪市上的回归结果和我们之前的总体回归结果基本相近，金融媒体情绪因子与投资组合的收益率呈现负向关系，即金融媒体的情绪越消极，该类投资组合的预期收益率越高。此外，从两个市场的对比效果来看，深市的定价误差明显小于沪市的定价误差，且沪市的四因子模型总体定价显著性更强，说明其更有可能拒绝$\alpha = 0$的原假设。另外，四因子模型在深市上的回归结果来看，金融媒体情绪因子的显著性也明显更高，这说明金融媒体情绪因子在深市的表现更好，结合之前深市相关个股的新闻咨询更多这一现象（几乎为两倍）来看，深市的金融市场相比沪市可能更加容易受情绪影响。

4.4.2 媒体情绪因子在不同行业中的定价能力

此外，本文不仅分市场做了回归分析，还在不同的行业中做了同样的回归分析。本文从 CSMAR 上下载了 A 股所有的行业分类以及对应的股票代码，一共有工业、公用事业、房地产、金融、商业、以及综合类物种行业，经过筛选以及对新闻的匹配后发现，综合类行业匹配到的新闻极少，大约为 3w 条，分布到 98 个月的整个 A 股市场上后，会有许多月份与个股公司无法对应相应的新闻，导致需要对变量进行大规模的补零。为了避免过多的缺失值影响实证效果，我们仅在前五个行业上进行回归分析，并考察其总体的模型定价误差，具体结果如下：

表 4-9　不同行业的回归分析

Model（6）	工业	房地产	金融	公用事业	商业
α	-0.0013	-0.0017*	-0.0135**	-0.087**	-0.0203**
	(-1.38)	(-1.83)	(-2.04)	(-2.38)	(-2.17)
R^2	0.436	0.401	0.372	0.351	0.365

注:*、**、***分别代表在 10%、5%、1%水平下显著相关

可以看出四因子模型在工业、房地产、金融、以及商业领域都取得了不错的定价效果，定价误差α基本为零。尤其在工业领域中，其定价误差已经基本不显著，基本维持了$\alpha = 0$的原假设，可以说，在工业领域上，引入金融情绪因子的四因子模型能够很好的解释该领域市场上超额收益产生的原因。而在房地产、金融与商业三个领域中，模型的定价误差α也极其接近于 0，不过三者分别在 10%、5%以及 5%的水平下显著，说明在这三种市场上，虽然四因子模型能够很好的解释一部分收益因素，但是仍然存在一部分增量信息是模型未能考虑到的。最后，模型在公共事业领域的表现不佳，其定价α并不接近于 0，且其定价误差的显著性也是五个领域中最高的，这可能是因为公共事业在国内具有较强的政府垄断性质，其一般由政府主导，收益较为稳定与客观，更容易受市场大环境因素的影响，而非相关的媒体情绪。

由此可见，我们的四因子模型在，除了公共事业这类国家垄断的行业以外，其余行业均取得了不错的效果，说明我们的结果是稳健的。由此可以认为，本文通过 Transformer 架构在金融语料库上训练出的 Fin-BERT 模型，能有效提取出新闻报道中的情感信息，并且在市场实证上得到了检验。这种情感的影响对市场的预期收益的是负向的，在遇到相关媒体报道时，投资者应当保持冷静，综合分析，避免随着报道而频繁地调整投资组合。

5.结论与展望

5.1 文章结论

本篇研究从金融市场上的非结构性数据入手，关注到资本市场上媒体情绪容易影响市场表现这一情况，拟从与个股相关的新闻出发，从中提取专业媒体对相应股票的情绪指标，以此构建金融媒体情绪因子。

为了在尽量准确的从新闻中提取媒体情绪，本文使用了自然语言处理（NLP）领域中目前最为汇报的 Transformer 结构，来训练本次研究相关的模型。同时，为了最大程度的降低数据获取以及标注的成本，我们还引入了自监督学习的框架（SSL），在两者结合的基础上，预训练了自己的 BERT 模型。在训练过程中，我们选择了 CSMAR 上的全部金融无监督新闻语料，并且采用了针对中文 BERT 预训练模型的方法——Whole-Word-Mask。在得到预训练模型的基础上，我们以目前领域内开源的三个中文 BERT 语言模型作为基准，在接入不同下游分类以及不同的新闻信息截断长度的情况下，进行了 36 组对比实验，结果发现通过本文所设想的模型下，训练出来的模型，在预训练模型维度、下游任务维度以及文本截断长度维度下的效果均是最优的，这说明，在有限成本得到控制的情况下，使用 Transformer 架构搭配自监督学习框架，就已经能够在现有的自然语言处理领域中得到针对金融研究的非常可观的效果，甚至在下游分类器中，我们无需接入复杂的网络结构，反而能够降低模型的过拟合程度，即降低运算成本的同时，还提高了模型的精度。

另外，我们还发现，针对不同新闻文本的截断长度，在模型精度与运算成本的平衡下，我们也有一定的选择空间。首先，我们发现，输入全部新闻文本长度的信息，模型的精度毫无疑问是最高的，但是在这种情况下，我们的运算成本以及数据存储成本将会是指数级的

上升，这就失去了我们在前面引入这个框架的初衷——即如何在数据要求较低、模型训练成本较小的情况下，提升模型的精度，从而服务到金融领域的实证分析以及量化投资。进一步，我们发现，大部分新闻文本由于其高度的规范性，通常在第一小段就会将新闻的信息进行高度总结概括，因此，我们也选择了文本截断长度为 128 的情况，惊喜的发现最终模型的分类精度仅下降了 0.1%左右，但是其训练速度却提高了十倍，特征数据的体量也见笑了数倍不止。然后进一步的缩短截断长度后发现，当将字数限定为接近一个摘要的字数时，由于官方新闻本身在开篇带有一些冗余信息，导致输入给模型的数据噪声极强，导致最终的模型精度大幅下降。

由此我们可以得出第一个结论，在金融领域的实证研究以及量化分析中，常常需要用到文本等非结构性数据。而使用传统的词频统计等方法容易割裂文本的语义从而使得模型的精度不高，从而影响后续的分析效果；但如果使用复杂的深度学习模型，往往对数据获取的成以及算力的资源要求极高。本文发现，通过将自然语言处理领域中最强大的 Transformer 模型框架与自监督学习（SSL）的框架相结合，使用 BERT 模型在金融场景下进行预训练，那么在回归到具体的下游金融任务中进行微调时，所需要的数据成本将会大幅度的减小。此外在选择基准的 BERT 模型重新进行预训练时，通过我们的研究界结论可以发现，尽量选择特征维度最大的模型，特征维度的差异将会导致模型精度的巨大差别。

通过我们的模型计算出了金融媒体的情绪分类，并通过对积极情绪、消极情绪、以及中性情绪的分类概率，计算出了最终的情绪指标。随后我们依照 Fama-French 三因子模型的研究范式，划分投资组合的方式，并依据中国市场上的三因子模型研究，我们构建了市场因子 MKT、市值因子 SMB、市盈率因子 PMN，通过 Fama-Macbeth 在整个 A 股市场回归发现金融媒体情绪因子确确实实能够影响 A 股市场上投资组合的超额收益,这种影响是负向的，即情绪越消极的股票或者投资组合，其预期获得的超额收益越大。通过 5×5分组发现，规

模较小、负面情绪较强的投资组合往往这种影响越明显，而投资组合的新闻报道偏中性时，对预期收益的影响不再显著。这说明，在中长期范围内，当遇到相关负面新闻时，投资者应尽量冷静，避免过多受到新闻报道情绪的干扰；另一方面，监管部门应当规范市场上的新闻报道，使其在情感上尽量偏中性。

此外，与前人提出的中国版三因子相比，我们的构建的金融媒体情绪因子在有额外的信息增量，包含了中国版三因子所没有的定价信息。最后，我们分别在沪市和深市上做了稳健性检验，发现两个市场都收金融新闻媒体的影响，并且其影响方向也都相同，不过相比于沪市，深市对金融媒体的情绪反应更加激烈，其市场上涉及的个股新闻数量也更多。另外一个实验是在工业、房地产、公共事业、金融以及商业上分行业进行了四因子模型的回归，我们发现，金融媒体情绪因子在工业、房地产、金融以及商业领域都有不错的定价效果，但是在公共事业领域其效果并不显著，究其原因，可能是因为公共事业的国家垄断性质，导致其受市场媒体情绪的影响较小。

总体而言，区别于以往的金融文本分析方法，本文的第一个创新点在于将 Transformer 框架引入金融文本分析中，并使用自监督学习算法框架控制数据分析的总体成本，在充分实验的基础上获取了成本可控的文本分析模型。其次，本文用该模型构造了新的定价因子——金融媒体情绪因子 PMN，并证明了该因子对超额收益有一定的解释能力，且结果稳健。

5.2 研究展望

本文通过结合了 Transformer 框架与自监督学习框架的 BERT 模型，在中文金融新闻文本上完成了模型的预训练与微调，并构造金融媒体情绪因子进行实证分析检验，得到了不错的效果。但是，本文在处理金融新闻时，首先于个人精力以及论文制作时长的限制，直至针对新闻和个股进行了一一对应，但这其实是一种较为粗糙的处理方式，会遗漏掉非结构性数据中的许多信息。

例如，市场上发生的同一件事情，指向同一个股和公司，但是由于发送媒体以及发送时间的不同，这类新闻往往有不同的新闻 ID，这就导致在计算媒体情绪时，会将相同的事件影响反复叠加；但另一方面，如果有多家媒体报道，就证明该事件影响力大，不能简单当做一条新闻处理，而是可以引入合适的指标计算媒体关注度。第二，一篇新闻与个股及公司往往也不是一一对应的，可能一条新闻能够涵盖多个公司、整个行业甚至整个市场，但是由于数据本身的限制，该类信息往往不会有相应的字段进行保存，需要研究者另外建设模型从新闻中做主体以及主题的分类。第三，即使是一个事件只有一篇新闻来描述，一篇新闻也只对应一只股票，但是股票与股票之间，公司与公司之间往往依然存在联系。一家公司的新闻往往也会在供应链上引起极大的波动，例如一家上市的养殖企业，如果他的上游供应链，即饲料出售方的产品出现了一些负面新闻，但由于该上游企业并未上市，因此当我们但从新闻文本中获取信息时可能会漏掉一些重要内容。

针对上述三点，本文也对未来的研究方向提出一些展望：

第一，构建一个自然语言处理中的主题提取模型，多每一篇新闻输出相应的内容主题，当多篇新闻主题重复时，应该尝试从中提取额外信息。

第二，构建一个新闻主体提取模型，该模型能够从新闻文本中识别出一个或多个新闻的主题角色，根据提取的主体对新闻进行个股、行业、市场的匹配，可能会达到更好的效果。

第三，可以根据公司之间的供应链关系，制作图神经网络，充分挖掘公司与公司之间相关联的潜在信息，从而研究供应链上相近的公司企业，在企业运营、管理决策上相互之间有什么影响。

参考文献

[1] Airoldi，E.M.，E.A.Erosheva，S.E.Fienberg，C.Joutard，T.Love，and S.Shringarpure，“Reconceptualizing the Clasification of PNAS Articles”，Procedings of the National Academy of Sciences，2010，107 (49)，20899-20904.

[2] Antweiler，W.，and M.Z.Frank，“Is Al That Talk Just Noise?The Information Content of Internet Stock Mesage Boards”，The Journal of Finance，2004，59 (3)，1259-1294.

[3] Athey，S.，“Beyond Prediction:Using Big Data for Policy Problems”，Science，2017，355 (6324)，483-485.

[4] Athey，S.，D.Blei，R.Donnely，F.Ruiz，and T.Schmidt，“Estimating Heterogenous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data”，Working Paper，2018.

[5] Bachmann，R.，S.Elstner，and E.R.Sims，“Uncertainty and Economic Activity:Evidence from Busines Survey Data”，American Economic Journal:Macroeconomics，2013，5 (2)，217-249.

[6] Baker，S.R.，N.Bloom，and S.J.Davis，“Measuring Economic Policy Uncertainty”，The Quarterly Journal of Economics，2016，131 (4)，1593-1636.

[7] Baker，M.，and J.Wurgler，“Investor Sentiment and the Cros-Section of Stock Returns”，The Journal of Finance，2006，61 (4)，1645-1680.

[8] Bali，T.G.，S.J.Brown，and Y.Tang，“Is Economic Uncertainty Priced in the Cros-Section of Stock Returns?”，Journal of Financial Economics，2017，126 (3)，471-489.

[9] Bali，T.G.，and H.Zhou，“Risk，Uncertainty，and Expected Returns”，Journal of Financial and Quantitative Analysis，2016，51 (3)，707-735.

[10] OLIVEIRA N，CORTEZ P，AREAL N.Stock market sentiment lexicon acquisition using micro blogging data and statistical measures[J]. Decision Support Systems，2016，85:62-73.

[11] Ang, A., Hodrick, R. J. and Xing, Y. et al. The cross-section of volatility and expected returns [J]. The Journal of Finance, 2006, 61(1): 259-299.

[12] Bali T.G, Cakici N. Idiosyncratic volatility and the cross section of expected returns[J]. Journal of Financial and Quantitative Analysis, 2008, (1)：29-58.

[13] Xu Y.,Malkiel B.G. Idiosyncratic Risk and Security Returns[J]. Ssrn Electronic Journal,2001.

[14] Goetzmann W. and Kumar A. Why do individual investors hold under -diversified portfolios? Unpublished working paper,2004,Yale University.

[15] Chng, M. T., Fang, V. and Xiang, V. et al. Corporate hedging and the high idiosyncratic volatility low return puzzle [J]. International Review of Finance, 2017, 17(3): 395-425.

[16] Switzer, L.N., Tahaoglu, C. and Zhao, Y. Volatility measures as predictors of extreme returns [J]. Review of Financial Economics, 2017, 35(1): 1-10.

[17] Bali, Turan G., Robert F. Engle, and Scott Murray. “Empirical Asset Pricing: The Cross Section of Stock Returns.” John Wiley & Sons, 2016.

[18] Bates, Charles E., and Halbert White. “Determination of Estimators with Minimum Asymptotic Covariance Matrices.” Econometric Theory 9.4 (1993): 633-648.

[19] Chincarini, Ludwig B. “Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management.” McGraw-Hill, 2006.

[20] Fama, Eugene F., and James D. MacBeth. “Risk, Return, and Equilibrium: Empirical Tests.” Journal of Political Economy 81.3 (1973): 607-636.

[21] Gallant, A. Ronald, and George Tauchen. “Which Moments to Match?.” Econometric Theory

12.4 (1996): 657-681.
[22] Gibbons, Michael R., Stephen A. Ross, and Jay Shanken. “A Test of the Efficiency of a Given Portfolio.” Econometrica: Journal of the Econometric Society (1989): 1121-1152.
[23] Giglio, Stefano, Yuan Liao, and Dacheng Xiu. “Thousands of Alpha Tests.” Chicago Booth Research Paper 18-09(2018).
[24] Hansen, Lars Peter. “Large Sample Properties of Generalized Method of Moments Estimators.” Econometrica : Journal of the Econometric Society (1982): 1029-1054.
[25] Harvey, Campbell R., Yan Liu, and Heqing Zhu. “… and the Cross-section of Expected Returns.” Review of Financial Studies 29.1 (2016): 5-68.
[26] Shanken, Jay. “On the Estimation of Beta-pricing Models.” Review of Financial Studies 5.1 (1992): 1-33.
[27] Wooldridge, Jeffrey M. “Estimation and Inference for Dependent Processes.” Handbook of Econometrics 4 (1994):2639-2738.
[28] Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. Journal of Financial Economics, Vol. 116(1), 1 – 22.
[29] Lee, C. M. C., Y. Qu, and T. Shen. Reverse mergers, shell value, and regulation risk in Chinese equity markets. Working paper. 2017.
[30] Zhou, Guofu. Lecture Notes on Empirical Asset Pricing. Southwestern University of Finance and Economics, 2019.
[31] Liu, J., R. F. Stambaugh, and Y. Yuan. Size and Value in China. Journal of Financial Economics, 2019.
[32] Tomas Mikolov, Kai Chen, Greg Cor- rado, and Jeffrey Dean. Efficient estimation of word representations in vector space. Computer Science, 2013.
[33] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, 2014.
[34] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.
[35] Ashish Vaswani, Noam Shazeer and Niki Parmar. Attention Is All You Need. Conference on Neural Information Processing Systems, 2017.
[36] Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
[37] Griffin, John M, Hirschey, et al. How Important Is the Financial Media in Global Markets?[J]. CFA Digest, 2012,42(2):51-53.
[38] Dawson P, Downward P, Mills T C. Olympic news and attitudes towards the Olympics: a compositional time-series analysis of how sentiment is affected by events[J]. Journal of Applied Statistics, 2014, 41(6): 1307-1314.
[39] Wei Y-C, Lu Y-C, Chen J-N, et al. Informativeness of the market news sentiment in the Taiwan stock market[J]. The North American Journal of Economics, 2017, 39: 158-181.
[40] Tetlock P C. Giving content to investor sentiment: The role of media in the stock market[J]. The Journal of finance, 2007, 62(3): 1139-1168.
[41] Seok S I, Cho H, Ryu D. Firm-specific investor sentiment and daily stock return [J]. The North American Journal of Economics and Finance, 2019, 50: 100857.
[42] Niu H, Lu Y, Wang W. Does investor sentiment differently affect stocks in different sectors? Evidence from China[J]. International Journal of Emerging Markets, 2021, ahead-of-print(ahead-of-print)
[43] Agarwal S, Kumar S, Goel U. Social media and the stock markets: an emerging market perspective[J]. Journal of Business Economics Management, 2021, 22(6): 1614-1632.
[44] He Z F, He L J, Wen F H. Risk Compensation and Market Returns: The Role of Investor Sentiment in the Stock Market[J]. Emerg Mark Financ Trade, 2019, 55(3): 704-718.

[45] Ni Z-X, Wang D-Z, Xue W-J. Investor Sentiment and Its Nonlinear Effect on Stock Returns--New Evidence from the Chinese Stock Market Based on Panel Quantile Regression Model [J]. Economic Modelling, 2015, 50: 266-274.
[46] Fisher K L, Statman M. Blowing Bubbles[J]. Journal of Psychology and Financial Markets, 2002, 3(1): 53-65.Kartick Gupta, Rajabrata Banerjee,
[47] Kartick G, Rajabrata B. Does OPEC news sentiment influence stock returns of energy firms in the United States?[J] Energy Economics, 2019, 34-45,
[48] 游家兴, 吴静. 沉默的螺旋:媒体情绪与资产误定价[J]. 经济研究, 2012, 47(07): 141-152.
[49] 姚加权, 张锟澎, 罗平. 金融学文本大数据挖掘方法与研究进展[J]. 经济学动态, 2020(4):16.
[50] 黄俊, 郭照蕊. 新闻媒体报道与资本市场定价效率——基于股价同步性的分析[J]. 管理世界, 2014(5):10.
[51] 郦金梁, 何诚颖, 廖旦,等. 舆论影响力、有限关注与过度反应[J]. 经济研究, 2018, 53(3):16.
[52] 李志生, 李好, 刘淳,等. 天使还是魔鬼?——分析师媒体荐股的市场效应[J]. 管理科学学报, 2017, 20(5):16.
[53] 马黎珺, 伊志宏, 张澈. 廉价交谈还是言之有据?——分析师报告文本的信息含量研究[J]. 管理世界, 2019, 35(7):19.
[54] 潘越, 戴亦一, 林超群. 信息不透明,分析师关注与个股暴跌风险[J]. 2021(2011-9):138-151.
[55] 权小锋, 吴世农, 尹洪英. 企业社会责任与股价崩盘风险:"价值利器"或"自利工具"?[J]. 经济研究, 2015, 50(11):16.
[56] 谭松涛, 甘顺利, 阚铄. 媒体报道能够降低分析师预测偏差吗?[J]. 金融研究, 2015(5):15.
[57] 唐国豪, 姜富伟, 张定胜. 金融市场文本情绪研究进展[J]. 经济学动态, 2016(11):11.
[58] 于琴, 张兵. 股市的媒体强化效应：“强者恒强”还是“盛极而衰” [J]. 山西财经大学学报, 2020, 42(06): 45-58.
[59] 王晓丹, 尚维, 汪寿阳. 互联网新闻媒体报道对我国股市的影响分析[J]. 系统工程理论与实践, 2019, 39(12): 3038-3047.
[60] 陈鹏程. 媒体情绪与 IPO 市场绩效:理论与实证[J]. 系统工程, 2017, 35(05): 37-43
[61] 黄宏斌, 刘树海, 赵富强. 媒体情绪能够影响投资者情绪吗——基于新兴市场门槛效应的研究[J]. 山西财经大学学报, 2017, 39(12): 29-44.
[62] 原思雨, 任龙. 媒体情绪传染与股票收益率[J]. 兰州财经大学学报, 2022,3 8(06): 82-97.
[63] 徐巍, 陈冬华. 自媒体披露的信息作用——来自新浪微博的实证证据[J]. 金融研究, 2016(3):17.
[64] 徐永新, 陈婵. 媒体荐股市场反应的动因分析[J]. 管理世界, 2009(11):9.

经世济民
孜孜以求

西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

校址：四川成都温江柳台大道555号
电话：028-87092032　传真：028-87092632
邮编：611130
网址：http://www.swufe.edu.cn
Address：Liulin Campus (Main Campus): 555, Liutai Avenue,
Wenjiang District, Chengdu, Sichuan, P. R. China
Tel：86-28-87092032　Fax：86-28-87092632　Postcode：611130