南京大学

本 科 毕 业 论 文（设 计）

院　　系　工程管理学院

专　　业　金融工程

题　　目　基于LDA主题模型的投资者情绪对股价影响研究

年　　级　2016　学 号　161278021

学生姓名　蒙劢

指导教师　张科　职 称　讲师

提交日期　2020年5月28日

摘 要

投资者情绪一直是金融领域研究的热点，并且从短期看，股票市场波动受投资者情绪影响较大。我国 A 股市场的散户投资者比例较高，使得对众多散户投资者情绪的研究具备了现实意义。

本文基于行为金融学理论和自然语言处理技术，选取上证 50 指数 15 只成分股，以东方财富股吧投资者发帖和评论文本为研究对象，分析其表现的投资者情绪对股票超额收益率的影响，进一步的，探究了投资者情绪对股价影响的非对称效应和对超预期盈余的解释力度。针对帖子常常反应投资者抱怨，从而产生大量噪声的问题，采用 LDA 主题模型对帖子进行主题分类，提取出海量帖子中与股价相关的隐含信息。针对帖子长度普遍较短而无法准确训练主题模型的问题，采用帖子+评论聚合模式将同一帖子下的内容连接成长文本进行主题模型训练。利用百度 AI 情感倾向分析技术和 Python 中文情感分析包 snownlp，以句子为单位进行情感倾向分析，分析各主题帖子展现的投资者情绪。

研究结果表明：(1) 使用经主题分类的情绪指标能提取出海量股票文本中的价值信息，与股票超额收益率正相关；(2) 不同主题情绪对股票超额收益的影响程度和持续性不同；(3)主题情绪对股票超额收益率的影响存在非对称效应；(4) 主题情绪能有效解释公司超预期盈余。本文研究结论对股市投资者情绪研究具有一定意义和价值，为该领域研究提供了一种独特的切入视角。

关键词： 投资者情绪；主题模型；LDA；股价；帖子

Abstract

Investor sentiment has always been a hotspot in financial research. The fluctuation of stock market is greatly influenced by investor sentiment in the short term. Retail investors account for a large proportion in Chinese A share market, so the research of investor sentiment becomes meaningful.

This article studies the impact of posts in financial forum on stock price, and furthermore, studies the asymmetric effect of the impact of investor sentiment on stock price and whether investor sentiment can explain unexpected surplus, which is based on behavior finance theory and nature language process technique. Regarding the fact that the posts in financial forum always contain complaint of investors and then, produce noise in text analysis, we utilize LDA topic model to classify the posts in order to extract implicit information which is relevant to stock price. Regarding the problem that the length of posts in financial forum is short, resulting in inaccuracy in LDA topic model training, we use a aggregation pattern named post-comment to concatenate the short text. Taking a sentence as a unit, we make use of Baidu AI emotional tendency analysis technology and snownlp(a Python sentiment analysis package) to analyze invertor sentiment in posts of different topics.

The results prove that: (1) The sentiment indicators which are produced by topic classification can extract value information from massive stock text data, and there is a positive correlation between stock excess return and investor sentiment. (2) Sentiment of different topics has different impact extent and persistence on stock excess return. (3) There is asymmetric effect of the impact of investor sentiment on stock price. (4) Investor sentiment that is classified by topic can effectively explain unexpected surplus. The conclusion of this article is of certain significance to the current research of investor sentiment and provide a new perspective in this research field.

Keywords: investor sentiment; topic model; LDA; stock price; post

目录

1 引言

1.1 研究背景

1970 年，Eugene Fama[1]提出有效市场假说，在他看来，市场是有效的，资产内在价值的所有信息已经完全且迅速地体现在资产价格上，因此大部分投资者无法战胜市场。假说中包含三个假设前提：（1）有关公司或市场的信息是公开且透明的，不存在或很少存在未披露的内幕信息，且这些信息的获取是没有成本或几乎是没有成本的，所有信息以一种理想化的方式同时传播到所有市场参与者手中；（2）信息的发布时间呈随机分布，先发信息与后发信息无关，市场无法预知何时会产生一个新信息；（3）大部分投资者是理性的，他们参与市场交易是出于自身利润最大化的目的，同时运用自身知识理性评估资产价值并进行交易。但是，在实际市场中，信息在不同投资者之间的传递速度是不同的，且由于投资的时间尺度不同，不同投资者关注的信息也有所差异，这使得人们根本无法对同样的信息做出一致的反应。此外，大部分投资者的投资行为都是受到非理性支配的，这种非理性源自存在于人类大脑中根深蒂固的认知偏差。

Robert Shiller[2]在 1984 年第一次向世人展示了“噪音交易者”模型（Noise Trader Model）。在此模型中，聪明投资者依据股票基本面价值进行投资，而噪音交易者的存在造成了股票价格与公司的内在价值出现了偏离。价值投资背后的核心逻辑是股票的价格围绕着其内在价值波动。在一个理性程度高的市场，前者主导，价格围绕价值的波动不大；而在一个理性程度低的市场，后者主导，价格往往大幅偏离股票基本面显示的内在价值。

Fama 和 Shiller 对有效市场理论的观点是完全不同的，但 2013 年诺贝尔经济学奖却同时授予了这两位经济学家和与 Fama 持相同立场的 Lars Peter Hansen。瑞典皇家科学院颁发诺奖时指出，“这些看起来令人惊讶且矛盾的发现正是今年诺奖得主分析做出的工作。”三位教授的研究成果“奠定了我们如今对资产价格理解的基础”。为了说明他们的贡献，委员会指出，这三位学者的发现表明“市场价格的波动受到理性和人性行为共同影响。这是科学的胜利，是人类在探究市场终极真相道路上坚实的一步。”因此，对噪声交易者的研究显得格外重要。

相较于准入门槛较高的期货、债券市场，门槛低、回报高的股票市场脱颖而出，使得越来越多的人们选择加入到散户投资者这一行列，希望通过股市的高波动来获取收益。近年来，中国境内股票市场投资者数量不断增多，投资风格也更加多样化。为了更全面、准确地把握市场运行规律，学者们选取从直接到间接的各类股票市场变量进行深入研究，极大丰富了股票市场的研究手段和理论。然而，散户投资者数量的增多不可避免地导致市场风格切换速度加快，羊群效应等市场截面异象逐渐增多，各类题材概念被轮番炒作，常常出现股价与公司基本面脱钩的现象，市场规律难以准确把握。因此，近年来，以人的非理性行为为研究对象的行为金融学被频繁应用于股票市场研究中，对传统金融理论无法解释的异象给出了令人信服的解释。行为金融学的两大支柱是有限套利和心理学，采用一种新范式解释金融市场截面异象，特别是传统研究范式中不能解释的那一部分。广义上说，行为金融学通过假设代理人并非完全理性来对金融市场截面异象进行解释。狭义上说，行为金融学主要放松了“理性”的限制条件，在这一基础上研究投资者行为。具体来说，在一些行为金融学模型中，代理人由于认知偏差和过度自信等原因，导致决策理念发生偏差；而在另一些模型中，代理人确实考虑了“理性”框架下的贝叶斯理论，但因投资者非理性心理及行为的存在，导致其投资决策不被广泛接受。

无论采用传统分析思路还是行为金融提出的“非理性”假设，最终目的都是要分析导致股价变动的因素。相关研究表明，公司财报中的基本面信息会导致股价产生变动[3]，而由于时常进行线下调研与跟踪，券商分析师和财经记者等信息中介能够提前预测上市公司的基本面信息，因此，券商分析师研报和财经新闻等信息能帮助预测公司盈利情况[4]。那么相比于分析师等中介，个人投资者是否也会掌握一些公司的基本面信息呢？Bagnoli 等(1999)[5]将分析师网站 First Call 的预测信息与个人投资者预测网站 whispers 的预测信息做对比，发现使用 whispers 网站的信息预测期望收益更为准确。Chen 等(2014)[6]通过分析主流财经媒体的文章与评论，发现个人在社交媒体发布的信息可以有效预测股票收益和公司意外盈余，提出“群体智慧”的概念。可见，个人投资者在社交媒体发布的信息可能包含着一些提前获知的上司公司基本面信息，在人数足够庞大的前提下，可能形成“群体智慧”，影响股价走势。

1.2 研究意义

由中国证券投资者保护基金有限责任公司发布的《2019 年度全国股票市场投资者状况调查报告》[7]显示：截至 2019 年 12 月 31 日，全国股票投资者数量约达 1.6 亿人，较上年同期增长 9.04%，其中散户占比 99.76%。在一个散户投资者占绝大部分的市场中，针对散户投资行为的研究就变得不可忽视。相较于机构投资者，散户投资者体现出的噪声交易者特征更为明显，更符合行为金融学的研究范畴。而散户的投资行为又常常能体现在各类股吧论坛上，包括交流投资心得和发表个人看法，甚至是各种抱怨，这类信息对股市的影响主要体现在两方面：（1）由于网络平台信息发布门槛低，其中不乏一些有关股市的虚假信息通过网络平台进行传播，从而造成股市剧烈波动，打击投资者对证券市场的信心，给股票市场带来负面影响。（2）由于网络舆情的传播具有羊群效应，在股票舆情事件发生后，就极易出现一些非理性的负面舆论信息，从而影响到投资者情绪，造成股市震荡。综上所述，无论基于股票投资者结构还是股市健康发展角度而言，在我国特殊市场环境下对散户投资者情绪的研究都具备现实意义。

此外，互联网技术的飞速发展使得个人动态、评论信息爆炸式增长，越来越多的人热衷于在网上发表、交流对某件事的看法，股市投资者当然也不例外。各类投资论坛上每天都有成千上万的帖子，交流对后市走向的看法，网络帖子能够直接反映出投资者情绪，而帖子数量的急剧增长又使得如何有效利用这类投资者发帖和评论信息成为当前的迫切需求。特别地，中国资本市场的透明化程度较欧美发达市场仍存在一定差距，不少价值信息通过非官方渠道发布和传播，不同投资者掌握着不同的有关公司基本面的信息，投资者在论坛上的交流可能代表着一种群体智慧，有助于指引投资。当前，对投资者发帖表达的情绪的研究已不在少数，但多数研究表明，股吧论坛上发表的帖子更多的是投资者的抱怨，这类抱怨难以对股价产生实质性影响。因此，本文旨在提取出投资者发帖中的隐含信息，过滤掉大量诸如抱怨等与股价无关的噪音，以期寻求到最能影响股价走势的信息，探求投资者情绪与股价等指标的真实相关性，并提供一种预测股价的新思路。

1.3 研究内容

本文的研究内容主要分为四个部分，第一章为引言，包括本文研究主题的发展背景和现实作用，阐述本文研究主题的合理性和创新性，并确定本文的研究思路与方法。第二章为文献综述，对本文涉及到的金融领域相关知识、自然语言处理、情感分析技术进行了必要的回顾与梳理，尝试发现现有研究的不足。第三章详细介绍了本文所用模型的推导及参数估计过程，是理解本文模型的基础。第四章为实证分析，构建股票超额收益率与投资者情绪的回归方程，检验二者的相关程度，并进一步探究了非对称效应和投资者情绪对公司超预期盈余的解释力度。最后一章为全文的总结与展望。

本文的总体研究思路如图 1 所示。

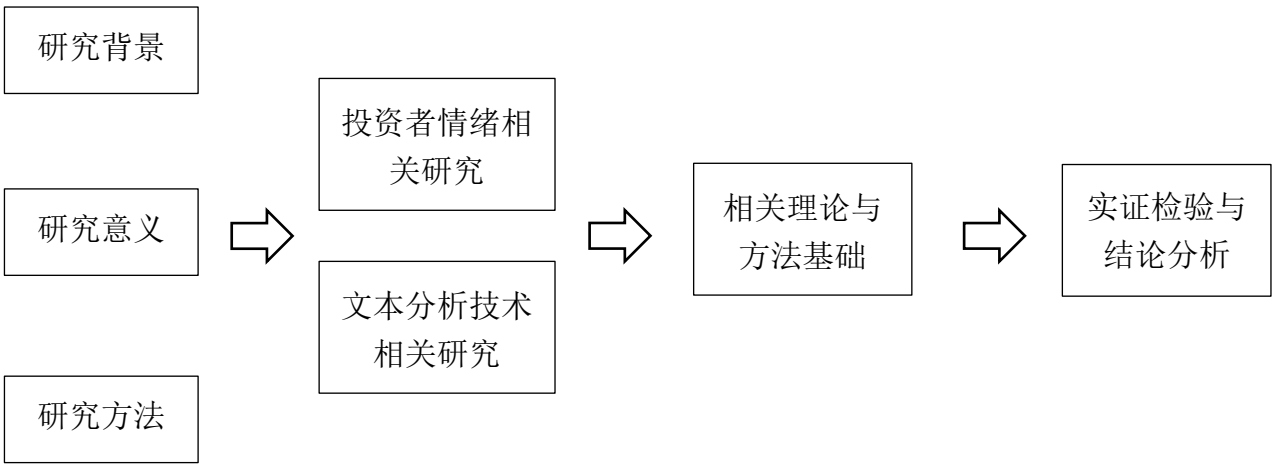

图 1 总体研究思路

1.4 研究方法

本文的研究过程主要分为三个部分，第一部分是建立基于 LDA 的主题分类模型，对文本数据进行分类，第二部分是对股吧论坛投资者发帖及评论的情绪进行量化分析，并构建相应的情绪指标，第三部分是将构建的情绪指标与股票超额收益率和公司盈余信息进行相关性分析。

在建立 LDA 模型方面，利用 Python 爬虫爬取来自东方财富股吧论坛的将近 200 万篇帖子，时间覆盖最早为 1993 年，直到 2020 年 2 月。为解决短文本训练

效果不佳的问题，将爬取到的数据按照“帖子+评论”的模式聚合为长文本，同时对文本数据进行去除停用词等的预处理工作。利用 Python 中文分词包 jieba 进行文本分词，将分词后形成的语料库传入 LDA 主题模型，通过 Python 机器学习库 scikit-learn 的参数优化工具，寻找 LDA 主题模型针对此语料库的最优主题个数，设置相关参数后代入建立的 LDA 模型进行训练，即可得到文本主题分类结果。

在对股吧论坛帖子情绪进行量化分析方面，借助百度 AI 情感倾向分析技术，以句子为单位进行情感分析，将句子得分加总得到整篇文档情感倾向。本文将投资者情绪分为三种类型——积极、中性和消极，分别对应得分 1、0 和-1。将得到的情感分析结果与主题分类结果结合，即可得到各主题的情感倾向，再以主题为单位构建相应的群体情绪指标，得到所有主题的情绪指标序列。

在相关性分析部分，添加一定数量的控制变量后，将经主题分类的情绪指标与股票超额收益率进行面板回归，区分实体效应和时间效应，再与直接使用未经主题分类的情绪指标进行回归的结果对比，以验证主题分类的有效性。最后，进一步探究了投资者情绪对股票超额收益率地非对称效应和对超预期盈余地解释力度。

1.5 创新之处

本文可能的贡献和创新点主要包括：第一，本文尝试将经典主题模型 LDA 应用至论坛文本分析中，丰富了利用股市信息分析投资者情绪的方法。目前针对中国资本市场散户投资者情绪的研究已不在少数，但在以散户为主的市场中，信息的交流与传递往往效率低下，充斥着投资者的抱怨情绪，少有研究将能够提取隐含信息的主题模型应用至投资者情绪分析中。第二，在分析投资者情绪与股票收益率相关性时，现有研究往往没有试图深究股票超额收益率的来源，由于公司的基本面信息是导致股价变动的重要因素，因此，本文在进行此类研究时，试图进一步研究投资者情绪与公司盈余信息的相关性，尝试对研究结论做更深层次的解释。

2 文献综述

2.1 投资者情绪相关研究

现有研究表明投资者情绪是影响股票价格的重要因素之一。早期，投资者情绪的度量由于受限于收集手段和情绪载体等因素，多采用金融市场指标、经济指数指标衡量投资者情绪，如Baker[8]、郑振龙[9]选取封闭式基金的折价率；Beer[10]、鹿坪[11]采用消费者信心指数等作为投资者情绪的代理变量，发现投资者情绪会对股票市场定价造成影响。相关研究虽然均表明了投资者情绪与股票市场价格走势之间存在一定相关性，但因其代理变量的间接性以及标准选取不统一致使研究成果分散，对研究结论间的横向或纵向比较造成一定困难。

近年来，大量金融信息通过网络发布和传播，学者开始以网络文本为研究对象，分析其与股票价格的相关性。这类研究主要以媒体新闻和论坛评论为研究对象，应用自然语言处理技术，提取与股价显著相关的信息。Lavrenko[12]将文本分析模型应用至金融新闻中，构建了一种将新闻情绪与股价走势相结合的预测系统，证实了根据这一系统推荐的金融新闻进行投资决策可以帮助投资者稳定盈利。Fung[13]建立了一个通过分析新闻文本来预测股价走势的系统，提出一种控制聚类法来筛选有效新闻。Antweiler[14]通过分析雅虎财经网站上与道琼斯工业指数中 45 只成分股相关的超过 150 万条评论数据，发现这一股票信息能显著预测股市波动性和股票收益率。Frisbee[15]分析了著名金融博主在其文章中表现出的乐观程度与标普 500 指数价格与成交量的相关性，发现当博主认为牛市即将到来时，股市会以更大概率进入熊市，反之亦然。造成这种现象的原因可能是，博客这一平台形式过于正式，不方便访问，导致信息传播速度缓慢。且博客仅代表博主个人观点，难以形成群体效应而被广泛认同，影响力度有限。而论坛帖子和评论因其易得性、公开性、广泛性等特点，更适合作为投资者情绪的研究对象。Wysocki[16]通过对 Yahoo 上超过 3000 只股票的分析，发现累计发帖量最多的股票往往具有异常收益率、高市值、高回报、高市净率、高波动和高成交量等特征。Tumarkin[17]通过对 72 支互联网行业股票的分析，发现在新闻消息异常活跃的日子，投资者观点的变化与行业调整后的异常回报相关。Das[18]将研究对象限定于 4 只股票，

构建新闻热度指标进行相关性分析，发现投资者情绪、新闻热度和股票价格三者高度相关。投资者情绪往往会受股市表现影响，却难以对股价产生实质性影响，说明投资者在股票市场中常常扮演接收者的身份，只能根据市场波动采取相应的策略。

随着机器学习技术的发展与计算机算力的提升，学者们开始将各种先进机器学习技术应用于股市文本的分析中，得出了更为精确的分析结论。李玉梅[19]分别采用三种文本分类和机器学习算法挖掘在股票论坛文本中包含的投资者情绪，再将投资者情绪分为看涨情绪和看跌情绪，并利用评论数量构建评论热度指标，以情绪指标和热度指标为解释变量，分析其对收益率等股票相关变量的影响。宋敏晶[20]利用支持向量机将投资者情绪分为看涨、看跌和中立情绪，从四个维度，即大盘表现、板块表现、行业表现和个股表现，分析投资者情绪对股票市场的影响。结果表明，从对大盘表现的影响看，大盘表现与投资者情绪存在因果关系，受投资者情绪影响。从对板块表现的影响看，投资者对某一板块的态度越乐观，这一板块越有可能获得超额收益。从对行业表现的影响看，食品饮料行业超额收益与投资者情绪正相关。从对个股表现的影响看，投资者情绪可以预测个股超额收益。莫倩等[21]结合基于模式和篇章结构的股评观点倾向性评分方法，将投资者情绪分为积极、消极和中性三种类型，利用标题和正文两种分类器来分别计算股评属于看多、看空和看平类别的可信度，研究结论表明结合两种评分方法进行股评分类具有较高的准确度。杨伟杰等[22]基于股评文章术语较多、情感特征词以动词为主和干扰信息多等特点，采用划分话语片段的模式查找主动词，将有倾向性的主动词保留为意见目标句，再将意见目标句进行情感分析，但由于缺少特定领域情感词典，分类准确度不高。张对等[23]为了探究股票市场投资者决策和股市表现是否受网络帖子和评论的影响，通过收集相关的网络股评并提取情感指数，建立 ARMAX-GARCH 模型，分析股评中情感因素与股市走势之间的相关性，对投资机构和个人的投资决策有一定的帮助。

2.2 文本分析技术相关研究

对文本语料库和其它离散数据集合建模的目标是找到对集合成员的简短描述，以便对大型集合进行有效的处理，同时保留基本的统计关系，这些统计关系

对分类、异常检测、摘要和相关性判断等任务都非常有用。

Baeza-Yates 和 Ribeiro-Neto[24]提出一种成功应用于现代互联网搜索引擎的方法，将语料库中的每个文档简化为实数向量，每个实数向量表示词汇的计数比率。Salton 和 McGill[25]提出 tf-idf 模式，对于文集中的每个文档选择“词”或“术语”作为基本单位，对每个单词的出现次数进行计数。经过适当的标准化之后，将这个词频计数与反向文档频率计数进行比较，后者度量一个词在整个语料库中出现的次数(通常以对数形式表达，然后再次进行适当的标准化)。最终结果是一个文档单词的矩阵 X，其列包含语料库中每个文档的 tf-idf 值，因此 tf-idf 模式可以将任意长度的文档缩减为固定长度的数字列表。虽然 tf-idf 模式在对集合中文档词集的基本识别中有一定优势，但是该方法并没有减少对文档的描述长度，而且会导致文档内部或文档之间的统计结构丢失。针对这些问题，学者们提出了几个其他的降维技术，其中最著名的是 Deerwester 等[26]提出的隐语义索引(LSI)。LSI 基于奇异值分解（SVD）的方法来得到文本主题，在大型语料库的压缩方面表现优异。但 LSI 本身存在一些问题，如 SVD 计算非常的耗时，尤其处理大型语料库时，对于这样的高维度矩阵做奇异值分解是非常难的，并且 LSI 不是一个完整的概率生成模型，其结果可能存在负值，而这样的负数结果往往难以解释。

为了研究 LSI 的相对优势和劣势，学者们开始开发文本语料库的生成概率模型，并研究 LSI 从数据中恢复生成模型的能力(Papadimitriou 等)[27]。Hofmann[28]提出了概率 LSI (pLSI)模型，也称为特征模型（aspect model），作为 LSI 的替代品。pLSI 用隐变量表示主题，应用概率生成模型对语料库进行主题分析，属于无监督学习方法。模型将文本表示为一个主题分布，将主题表示为一个单词分布，文本的生成过程为：首先，假定主题本身的概率分布；然后，在给定主题条件下计算文本和单词的条件概率分布。概率 LSI 的核心思想是寻找文本包含的潜在主题，但 pLSI 只提供了主题层面的概率生成模型，而没有提供在文档层面上的概率模型，使得必须在确定文本语料库的情况下才能对模型进行随机抽样，这就导致了几个问题：（1）随着文本语料库规模和主题个数的增加，需要估计的参数数量随之线性增长，参数数量过多可能导致过拟合；（2）当获得一篇新文档时，没有科学的方法为其分配一个概率分布。

为解决 pLSI 存在的问题，Blei 等[29]在 pLSA 的基础上引入一个文档层面的

狄利克雷先验分布，提出 LDA 模型。LDA 使用完整的概率统计方法进行文本建模，并基于词袋模型假设，即可以忽略文档中单词的顺序，更一般地，可以忽略语料库中文档的顺序。因此，LDA 没有把重心放在语法规则细节上，而是基于词频分布等统计学方法挖掘文本单词的结构和相关性，更适用于语法不规范、口语化的网络帖子文本分析。其次，LDA 将文本表示为主题的概率分布，将主题表示为单词的概率分布，克服了 pLSA 对文本主题概率分布的限制，更符合现实语境。概率分布的表达方式能够提取文本中的隐含信息，减少无关噪声的干扰，从而更精准的量化文本信息。

LDA 与 pLSI 同属于无监督学习方法，且提供了文档和主题层面的概率生成模型，已成为目前主流的文本建模方法。但 Lu 等[30]提出，LDA 模型的训练效果与文档长度密切相关，短本文中的单词由于缺乏足够的出现次数，其相关性判断将变得很困难。鉴于此，Hong 等[31]在针对社交媒体 Twitter 的文本分析中提出采用用户模式和术语模式分步训练 LDA。用户模式是指将同一发帖人的帖子聚集成一个长文本，术语模型是将具有同一术语的帖子聚集成长文本，这在一定程度解决了网络本文长度较短的问题，但同时建模过程需要大量的文本预处理和清洗工作。为了获得更好的训练效果，可以考虑引入新信息协助 LDA 模型训练。Michal[32]将文章作者信息与文本信息相结合，提出 ATM(author-topic model)模型，适用于文学作品的建模分析。由于引入了作者信息，ATM 使用“作者－主题”分布取代传统 LDA 的“文本－主题”分布，因此，该模型要求训练文本具有比较鲜明的作者特征。

LDA 提供了文档和主题层面的概率生成模型，由于其在文档层面引入了一个先验概率分布，需要估计的参数数量不随语料库中文档数量的增长而增加，比较适用于大型语料库的建模分析。由于情感分析领域通常针对大型语料库进行研究，因此 LDA 模型的应用频率不断提高。

Titov 等[33]提出了一种从在线用户评论中提取评价对象特征的多粒度 LDA 模型(multi grain-LDA，MG-LDA)，同时采用聚类算法，可以提取被用户评价的那一部分主题，而不是像传统主题模型那样对某一对象的全局特性进行评价。基于 MG-LDA，Titov 等[34]进一步提出了评论文本情感分析模型，旨在从用户对产品的评价和打分中提取情感观点。虽然 Titov 等通过引入聚类模型提取出高细粒

度的主题，但是模型中需要使用带有用户打分标签的评论信息，无法应用在更一般画的文本情感分析中，仅适用于对产品评论进行建模分析，应用范围具有一定的局限性。Zhao 等[35]同样针对用户在线评论情感分析问题，提出 ME-LDA 混合模型(MaxEnt-LDA)，该模型可以同时识别主题和主题观点词，但与 MG-LDA 一样，都需要监督学习，缺乏领域移植性。

为使主题模型兼具主题高细粒度和无监督学习等特点，学者们不断对主题模型进行改进。Brody 等[36]选择以句子为单位形成文档，建立“句子—主题—词”关系。由于 LDA 模型基于词袋假设，模型本身已经忽略了文档之间的结构关系，将一个句子作为一篇文档则进一步忽略了句子之间的结构关系，可能导致得到的主题词相关性较差，脱离了基本的语义学基础。另外，该方法将情感词的识别对象局限于提取出的主题词，而没有生成句子层面及文档层面的情感分布，不属于严格意义上的情感生成模型。Jo 等[37]提出 ASUM 模型，将主题和情感分析的采样单位设定为一个句子，即认为一个句子只能属于一个主题和一种情感，但从语义学角度来看，一个句子包括很多单词，这些单词可以分属于不同主题和不同情感，以句子为单位采样的方法同样忽略了词之间的结构关系，分类效果不佳。Lin 等[38]提出将主题情感联合的 JST 模型，属于无监督学习，证实其模型应用在电影评论的观点提取和情感分析上的效果较好。

一般而言，主题情感混合模型在语义学背景下通常采用两种处理方式。第一种是认为主题词和情感词不存在冲突，一个词可以既是主题词又是情感词，如 ASUM 模型和 JST 模型。另一种是将情感词与主题词分开，一个词或属于情感词，或属于主题词，分别处理，如 Mei 等[39]提出的 TSM 模型。TSM 模型利用改进后的 HMM 模型分别训练主题模型和情感模型，虽然研究结果表明 TSM 模型在分析网络博客时的表现较好，且具备一定的移植性，但将主题词和情感词分开建模的做法不太符合传统语义学背景，主题词不应与情感词硬性分开。针对这一不足，孙艳等[40]提出一种同时采样主题词和情感词的主题情感混合模型(UTSU 模型)，属于无监督学习方法。UTSU 假设一个句子表达一种情感，句子中的不同词语可以分属于不同主题，这一做法继承了传统 LDA 模型采样主题词的做法，同时更符合语义学背景，将主题词与情感词采样完美统一起来。UTSU 存在的不足是只考虑了正面和负面两种情感，而没有将情感量化。

3 文本建模理论与方法基础

3.1 LDA 模型

在介绍 LDA 模型之前，需要先介绍几个重要的概率分布和有关分布共轭的知识，这些知识是理解 LDA 主题模型的基础。

3.1.1 Beta 分布

服从参数α和β的 Beta 分布的概率密度函数为：

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$$
$$=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$
$$=\frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

其中，$\Gamma(z)$为Γ函数，函数式为$\Gamma(z)=\int_0^\infty \frac{t^{z-1}}{\mathrm{e}^t}dt$，$\mathrm{B}(x,y)$为B函数，函数式为$\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt$。

3.1.2 Multinomial 分布

假设一个人要从一个袋子中取出n个不同颜色的球，在每次抽取后拿掉这个球。相同颜色的球是等价的。X_i表示取出球的颜色为$i(i=1,\dots,k)$，p_i表示取到颜色为i的球的概率，则这个多项分布的概率密度函数为：

$$f(x_1,\dots,x_k;n,p_1,\dots,p_k)=\Pr(X_1=x_1 \text{ and } \dots \text{ and } X_k=x_k)$$
$$=\begin{cases}\dfrac{n!}{x_1!\cdots x_k!}p_1^{x_1}\times\cdots\times p_k^{x_k}, & \text{when } \sum_{i=1}^{k}x_i=n\\ 0 & \text{otherwise}\end{cases}$$

可以用Γ函数表示为：

$$f(x_1,\dots,x_k;p_1,\dots,p_k)=\frac{\Gamma(\sum_i x_i+1)}{\prod_i\Gamma(x_i+1)}\prod_{i=1}^{k}p_i^{x_i}$$

3.1.3 Dirichlet 分布

参数$\alpha_1,\dots,\alpha_K>0$，维度$K\geq 2$的狄利克雷分布，在基于欧几里得空间$R^{K-1}$里的勒贝格测度有个概率密度函数$Dir(\vec{\alpha})$，定义为：

$$f(x_1,\dots,x_K;\alpha_1,\dots,\alpha_K)=\frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

其中$x_1,\dots,x_{K-1}>0$且$x_1+\cdots+x_{K-1}<1$，$x_K=1-x_1-\cdots-x_{K-1}$。归一化参数$\mathrm{B}(\alpha)$是多项B函数，可以用Γ函数表示：

$$\mathrm{B}(\alpha)=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)},\alpha=(\alpha_1,\dots,\alpha_K)$$

3.1.4 Dirichlet-Multinomial 共扼

贝叶斯统计学派通常会先给数据赋予一个先验分布，通过对数据的采样分析可以获得对数据的知识，结合先验分布即可得到数据的后验分布，具体流程如下式所示：

$$事先给定先验分布+从数据获得的知识=数据后验分布$$

若事先通过经验给定数据的先验分布为$Dir(\vec{p}|\vec{k})$，再通过分析数据获得关于数据的多项分布$Mult(\vec{m}|\vec{p})$，则有

$$Dir(\vec{p}|\vec{k})+Mult(\vec{m}|\vec{p})=Dir(\vec{p}|\vec{k}+\vec{m})$$

以上式子实际上描述的就是 Dirichlet-Multinomial 共轭。这里的“共轭”指的是数据的后验分布与先验分布能保持一致性，不会因为从数据本身获得的知识而发生改变。具体而言，就是当给定先验分布为 Dirichlet 分布，若从数据获取的知识为 Multinomial 分布，则其后验分布也为 Dirichlet 分布，保留了参数的物理意义，方便对数据的真实分布做出合理解释。

3.1.5 LDA 生成过程

LDA（Latent Dirichlet Allocation）是一种语料库层面的文档和主题生成概率模型，它使用主题分布来表示文档，又用单词分布来表示主题。具体地，LDA 假

定按如下过程依次生成语料库D中的文档$\mathbf{w}$：

1、 选择N~Poisson(ξ)，即假设文档中词的数量N服从参数为ξ的泊松分布；

2、 选择$\theta \sim Dir(\alpha)$，即假设主题分布θ服从参数为α的狄利克雷分布；

3、 对于N个词中的词w_n：

(1) 选择一个主题z_n~Multinomial(θ)；

(2) 从$p(w_n|z_n,\beta)$（一个以z_n为主题的条件多项概率）中选择一个词w_n。即根据指派的主题所对应的词分布中采样出词w_n（每个主题有各自的词分布，先验分布是参数为β的 Dirichlet 分布，从单词数据中获取到多项分布的知识）。

重复上述过程，直到M篇文档都完成。

给定参数α和β，主题混合分布θ、主题$\mathbf{z}$（包含K个主题）、文档$\mathbf{w}$（包含N个词）的联合分布为：

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta)$$

其中，$p(z_n|\theta)$就是θ_i，对唯一的i满足$z_n^i=1$，表示当主题分布为θ的情况下，从某篇文档$\mathbf{w}$中采样得到词n的主题为z^i的概率，对于一个词汇w_n，若出现在主题z^i中，则$z_n^i=1$）。为得到文档的边际分布，首先以θ为积分变量进行积分，再对z的所有可能取值求和：

$$p(\mathbf{w}|\alpha,\beta)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

最后，将所有文档的边际分布相乘，就得到一个语料库的概率分布：

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

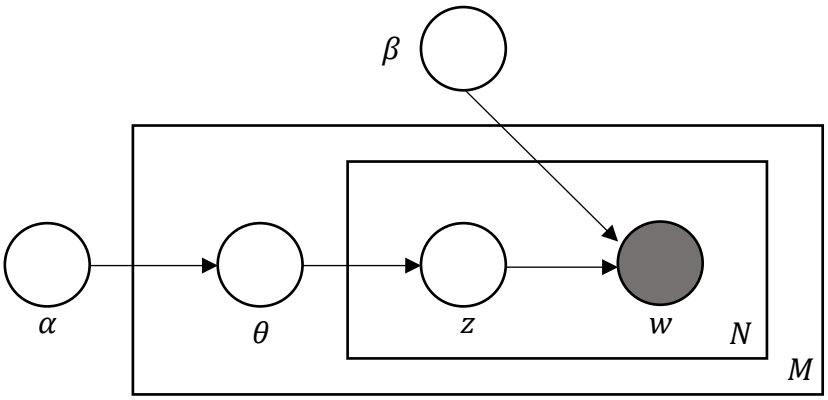

图 2 LDA 的图形表示[1]

LDA 的概率图模型如图 2 所示，LDA 模型参数可以分为三个不同的层面，语料库级参数α和β分别表示主题 Dirichlet 分布和词 Dirichlet 分布参数，通过事先抽样得到。文档级变量θ_d表示文档的主题分布，对一篇文档采样一次。字（词）级变量z_{dn}和w_{dn}分别表示文档主题和单词主题，采样对象设定为每篇文档中的每一个单词。

3.1.6 LDA 参数估计

为了使用 LDA，需要解决的关键问题是计算给定文档的隐含变量的后验分布：

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

但这一后验分布通常难以处理[41]，不过 LDA 可以应用各种近似推理算法，包括拉普拉斯近似、变分推断 EM 算法和 Gibbs Sampling，本文介绍最常用的变分推断 EM 算法。

3.1.6.1 变分推断 EM 算法

EM 算法基于 Jensen 不等式，从一个初始点出发，通过一种类似坐标上升法的方式优化对数似然函数，直到其达到一个局部极值（实践中要设定一个判断收

[1] 外框代表文档，内框代表文档中主题和词语的重复选择。

敛的条件），即似然函数的下界。EM 算法需要使用以变分参数为索引的下界族。通过一个特定的优化过程可以找到最可能的下界，同时决定相应的变分参数。由于 LDA 模型相对比较复杂，下界族通常难以获取，可以采用修改 LDA 图形模型的做法来简化似然函数的定义。

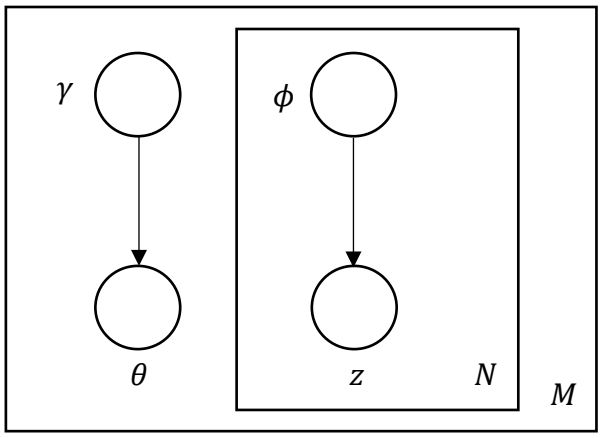

图 3 LDA 中用于近似后验的变分分布的图形模型表示

考虑图 2 中所示的 LDA 模型。θ和β之间的耦合是由于θ，z和w之间的边而产生的。这些边和w节点可以通过加入潜在变分参数γ和ϕ而大大简化（如图 3 所示），同时获取到变分参数的分布族，这个分布族具有如下所示的变分分布：

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n)$$

其中自由变分参数γ为狄利克雷参数，$(\phi_1, \ldots, \phi_N)$为多项参数。

上式所示为简化后的变分分布，求解这一分布的下界需要设定一个相应的优化过程[29]：

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z}|\gamma, \phi) \parallel p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$

因此，变分参数的最优值可以通过最小化变分分布和真实后验$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$之间的 Kullback-Leibler（KL）散度获得，迭代过程如下：

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i)|\gamma]\}$$

$$\gamma_i = \alpha_i + \Sigma_{n=1}^{N}\phi_{ni}$$

多项参数更新的期望计算如下：

$$\mathrm{E}_q[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\Sigma_{j=1}^{k}\gamma_j)$$

其中Ψ是 logΓ的一阶导数，可以通过泰勒近似计算。

变分分布实际上是一个条件分布，随 **w** 的变化而变化，可以将得到的变分分布写为$q(\theta, \mathbf{z}|\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$，这里显式地指出了参数 **w**。因此，变分分布可以看作是后验分布$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$的近似。在文本语言中，优化参数$(\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$因文档而异。特别地，可以将狄利克雷参数$\gamma^*(\mathbf{w})$看作是在主题层面提供了一篇文档的表达。LDA 变分推断的每次迭代都需要$O((N+1)k)$次运算。相关研究表明[29]，单个文档所需的迭代次数与文档中的单词数同阶，将产生大约N^2k阶的运算数量。

有了变分分布，可以推理出 EM 迭代算法步骤如下：

(1) （E-步）对每篇文档，找到变分参数$\{\gamma_d^*, \phi_d^*: d \in D\}$的最优值。如前文描述完成这一步骤。

(2) （M-步）对相应的模型参数α和β，最大化对数似然函数的下界。这相当于在 E-步计算的近似后验下为每个文档找到最大似然估计值和预期的充分统计。

条件多项参数β的 M 步更新解析式如下：

$$\beta_{ij} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j$$

重复 E-M 步骤，直到对数似然函数取到的最大化的下界。

3.1.7 LDA 模型优势

与混合 Unigram 模型相比，LDA 模型通过增加文档层面的 Dirichlet 分布参数，使得一篇文档可以表示为主题的分布。与 pLSI 模型相比，LDA 通过引入隐含变分参数，将参数估计过程与语料库中的训练文档剥离开来，从而克服了待估计的参数数量随着训练集文档的数量线性增长这个问题，同时也克服了 pLSI 可能存在的过拟合问题。

下面简要介绍在 LDA 出现之前的文本分析模型，以便能更清晰地认识 LDA 模型的优势。

3.2 其他文本分析模型

3.2.1 Unigram 模型

假设词典中一共有V个词$v_1, v_2, \ldots, v_V$，在 Unigram 模型下，每个文档的单词是由单个多项分布独立得出的：

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

模型结构如图 4 所示。

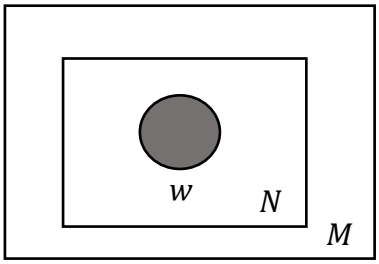

图 4 Unigram 模型

可见Unigram模型假设一篇文章只属于一个主题，这种假设往往不符合语义。

3.2.2 混合 Unigram 模型

Nigam 等[42]在 Unigram 模型基础上引入主题概念，提出混合一元模型（图 5）。模型中的主题z通过随机采样获得，文档词由以主题z为条件的条件多项式$p(w|z)$生成，从而生成一个文档（该文档中的所有词都来自一个主题）。一篇文档的概率分布为：

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z)$$

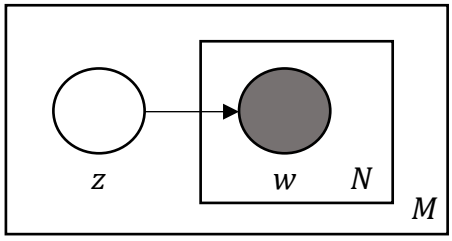

图 5 混合 Unigram 模型

虽然混合 Unigram 引入主题的概念，可以将主题表示为单词的分布，但模型知识假设一篇文档的主题数量不能超过 1。相关研究已经表明，这种假设通常限制性太强，无法有效地建模大量的文档[29]。

3.2.3 概率隐语义索引

概率隐语义索引(pLSI)进一步放宽了混合 Unigram 中一篇文档只包含一个主题的假设。如图 6 所示，给定未知的主题z，pLSI 模型假设文档标签d和单词w_n是条件独立的：

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d)$$

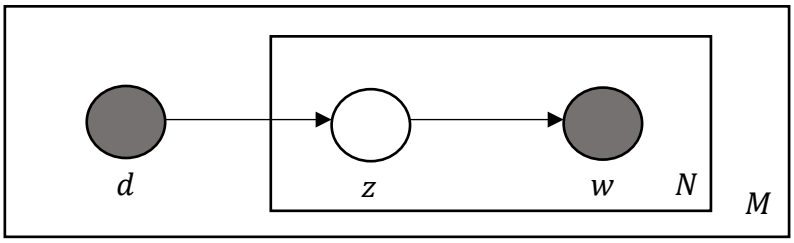

图 6 pLSI 模型

需要说明的是，pLSI 模型确实允许一篇文档包含多个主题，用$p(z|d)$表示文档中的主题分布。然而，d并不是针对整个语料库而言，而仅仅针对训练集文档。因此，d是一个 Multinomial 随机变量，输入模型的训练集文档数量越多，d的取值越大，模型只针对训练集文档的条件分布$p(z|d)$进行采样和学习。因此，pLSI 不能称为完整的文档生成模型，无法将概率分配给训练集之外的文档。

pLSI 的另一个困难是当训练集文档数量增长时，需要估计的参数数量随着随之线性增长，这可能导致模型出现过拟合的问题。Popescul 等[43]曾使用回火算法来避免发生过拟合，但研究结果证明，即使这样做也很难避免模型的过拟合问题。

4 实证分析

首先，本文选取的文本数据来源于东方财富网股吧，股票样本为随机选取的上证 50 指数成分股中的 15 只股票，对每只股票爬取自该只股票上市首日起的帖子，平均每只股票包含近 12 万的帖子数和近 20 万的评论数。通过中文分词、去除停用词、剔除无关数据等方式进行数据预处理。随后，运用 LDA 主题模型对所有爬取到的文本数据进行统一主题训练、分类。但有关研究[30]表明，当输入 LDA 模型的文本长度过短时，由于文档词频数较少，模型在词语相关性判别上容易出现偏差，模型训练效果受到影响。而在股票论坛上，同一帖子的内容和评论会存在一定程度的关联关系，即在同一篇帖子下的讨论可能只针对某一内容发表观点。因此针对帖子长度普遍较短、不利于 LDA 模型训练的问题，采用帖子＋评论聚合模式，将帖子及评论连接成长文本，作为一篇完整的文档传入 LDA 模型中进行训练，以文档为单位采样主题词，得到各主题的主题词以及各文档分属于某个主题的概率。

接下来，借助百度 AI 开放平台自然语言处理技术和 Python 中文自然语言处理模块 snownlp，以句子为单位对每篇文档进行情感分析。首先得到每个句子的情感倾向，将整篇文档的句子得分加总，得到一篇文档的情感倾向，再将当天所有文档的情感倾向加总，得到当天投资者的情感倾向。随后，结合 LDA 主题模型的分类结果，得到各主题下文档的情感倾向。汇总每天的情绪得分后，以股票超额收益率为因变量，各主题情感倾向为自变量，并添加一定数量的控制变量，建立回归模型，进一步地，探究了投资者情绪对股价影响的非对称效应和对超预期盈余的解释效果。通过回归分析，得到与股价相关性最强的主题情感及投资者情绪中包含的价值信息，并与直接用未经主题分类的情感倾向作为自变量的回归模型进行比较，分析加入主题模型的必要性，具体流程如图 7 所示。

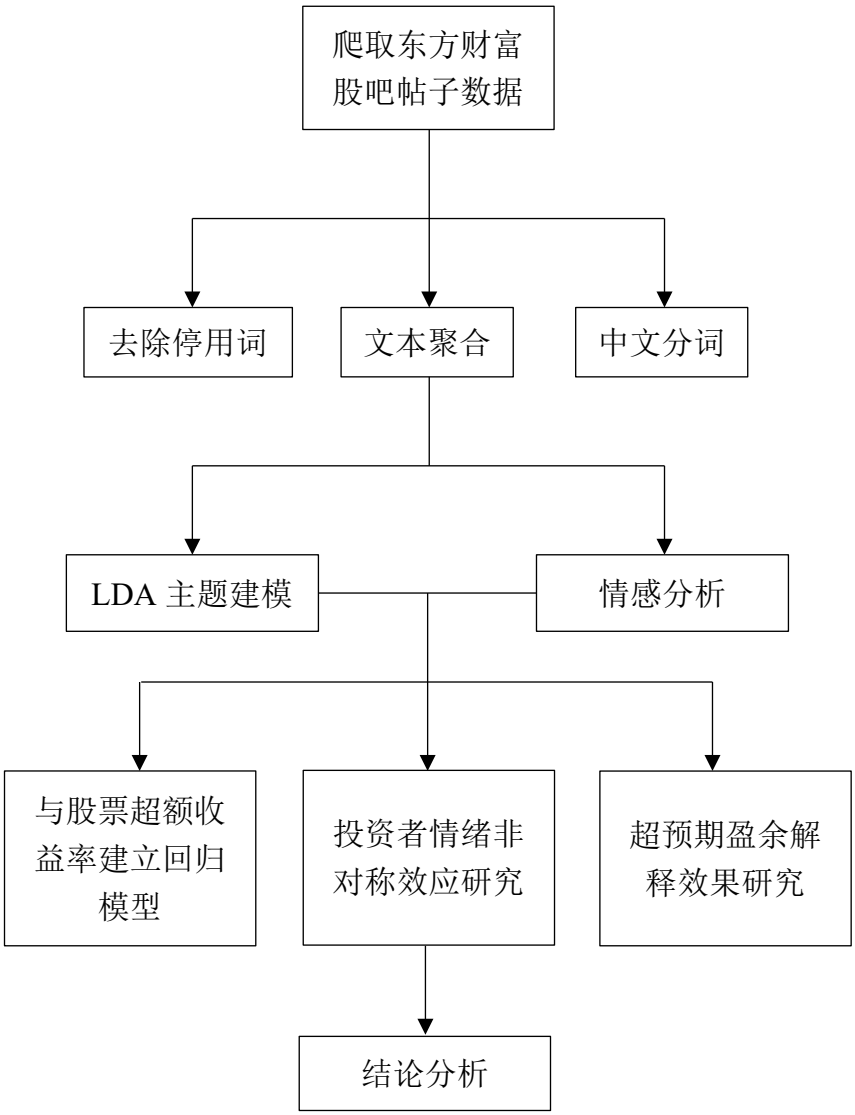

图 7 实证分析流程

4.1 股吧帖子文本数据采集

4.1.1 样本选择

本文选取的数据来源为东方财富网。该网站是国内影响力较大的财经媒体，全方位提供财经报道及金融市场信息，涵盖了证券、金融、基金、股票、债券、等领域。东方财富股吧论坛是投资者交流投资心得的社区。社区涵盖个股、主题、行业及概念四大板块，同时还设有访谈、悬赏问答、问董秘等模块，用户可通过多种形式进行互动。股吧论坛活跃度高，用户数量和每日发帖数庞大，信息来源丰富，各项网站搜索指标排名居同类网站前列，论坛还按照板块、行业、个股等

进行了详细的分类，方便分类采集数据。同时，网站可以获取自股票上市之日起的所有数据，适用于本课题的研究。

本文选取的数据样本为上证 50 指数的随机 15 只成分股。上证 50 指数成分股多为优质大型蓝筹股，从其分布情况看情况看，具有如下特征：第一，是优质蓝筹股的突出代表；第二，具有较好的流动性；第三，反映主流机构持仓的风向标。因此，上证 50 指数成分股的市场代表性强，适用于本课题的研究。

数据样本概览如表 1 所示。

表 1 样本概览

股票代码	贴子数	起始时间	结束时间
600016	124489	2014/2/27	2020/2/16
600309	96635	2001/4/21	2020/2/16
600340	113545	2006/5/9	2020/2/16
600690	135432	1993/10/15	2020/2/16
600703	159189	1999/9/3	2020/2/16
600887	189390	1997/1/9	2020/2/16
601138	95336	2015/8/11	2020/2/17
601166	223598	2008/5/10	2020/2/17
601186	203332	2010/3/17	2020/2/13
601211	82571	2014/6/13	2020/2/17
601236	59456	2011/12/7	2020/2/17
601288	229736	2010/6/16	2020/2/16
601319	59690	2018/6/6	2020/2/17
601328	111468	2007/5/15	2016/10/24
601336	31781	2011/11/21	2016/5/20

4.1.2 数据预处理

对于利用上述爬虫获取到的帖子文本数据，本文采用帖子＋评论聚合模式，将帖子与评论连接作为一个文档。由于 LDA 基于词频统计训练模型，因此需先将文本进行分词处理。本文采用 Python 经典中文分词模块 jieba 进行分词，再综

合利用哈工大停用词表、中文停用词表、百度停用词表和四川大学机器智能实验室停用词库，以及一些自定义的无关词语，删除对主题和情感判断没有帮助的词。由于有的帖子会发布上市公司公告、财经新闻和当日股价复盘等，这类帖子往往字数较多，但包含的投资者评论数据较少，因此，本文对原文字数超过 300 字的帖子予以剔除。最后再去除一些特殊符号及 html 标签，处理后的文本数据集如表 2 所示，词云如图 8 所示。

表 2 预处理后的数据集

股票代码	股票简称	剔除的无关贴子数	现有帖子数量
600016	民生银行	10823	113666
600309	万华化学	8968	87667
600340	华夏幸福	13158	100387
600690	海尔智家	10516	124916
600703	三安光电	9272	149917
600887	伊利股份	13665	175725
601138	工业富联	5735	89601
601166	兴业银行	20625	202973
601186	中国铁建	14528	188804
601211	国泰君安	11823	70748
601236	红塔证券	913	58543
601288	农业银行	5665	224071
601319	中国人保	541	59149
601328	交通银行	3464	108004
601336	新华保险	1733	30048

图 8 词云图

4.2 LDA 建模与情感分析

4.2.1 LDA 建模

本文参考 Blei(2003)[29]提出的方法，基于变分推断 EM 算法建立 LDA 模型。建立 LDA 模型需要确定一个重要参数——主题数量，Blei(2003)[29]采用基于困惑度的方法确定主题数，困惑度（Perplexity）的计算公式为：

$$Perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

其中，D表示语料库，M表示语料库中的文档数量，N_d表示每篇文档d中的单词数，w_d表示文档d中的词，$p(w_d)$表示文档中词w_d产生的概率。

根据困惑度的计算公式，更大的似然函数值将带来更小的困惑度，因此，参考 Griffiths(2004)[44]的做法，本文采用基于似然函数值的方法来确定最优主题数，似然函数值越大，模型拟合效果越好，不同主题数模型的似然函数值如图 9 所示。

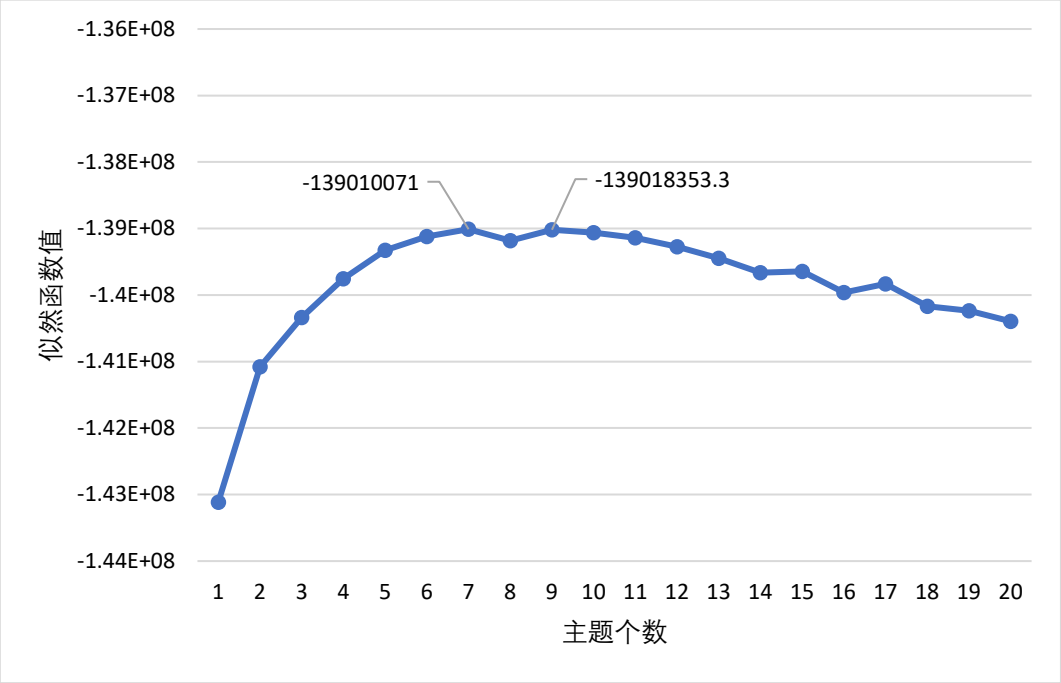

图 9 不同主题数的似然函数值

可见，似然函数值随主题数的增加先增大后减小，当主题数为 7 时，模型似然函数值最大，对应的困惑度为 4996.54。因此，本文选取的 LDA 模型主题数量为 7。另外，设置 EM 算法的最大迭代次数为 1000。主题训练得到的各主题前十大主题词如表 3 所示，可见，各主题词存在直观上的相关关系，表明主题分类结果较好。根据主题词之间的相关关系，分别给主题命名。各主题包含的帖子数量如图 10 所示，可见，股吧投资者讨论的话题多集中在大盘行情、板块行情和资金博弈上。

表 3 各主题十大主题词

主题名	公司发展	大盘行情	板块行情	行业发展	两融数据	资金博弈	公司红利
主题词	公司	银行	板块	基金	融资	主力	支持
	股份	股票	农行	市场	融券	涨停	业绩
	公告	大盘	伊利	资本	余额	筹码	股份
	产品	股市	海尔	资产	数据	跌停	每股
	项目	市场	垃圾	企业	偿还	散户	分红
	建设	股价	股票	增长	买入	资金	股价
	企业	资金	持有	公司	卖出	机构	增发
	中国	下跌	涨停	行业	证券	出货	集团
	发展	行情	指数	经济	交易	净流入	股东
	产业	投资	成本	未来	价值	成交量	董事会

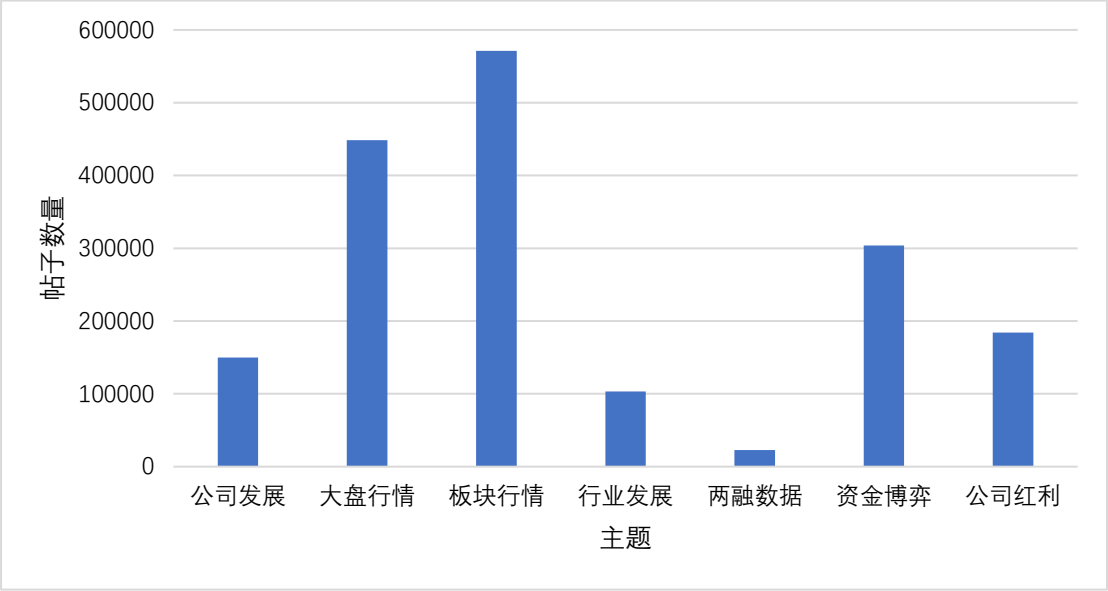

图 10 各主题包含的帖子数量

4.2.2 情感分析

本文借助百度 AI 情感倾向分析技术和 Python 经典中文情感分析包 snownlp，分析各帖子反映的投资者情绪。百度 AI 情感倾向分析可以对包含主观观点信息

的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。部分股票文本情感分析结果如图 11 所示，所有主题情感分析结果如图 12 所示。

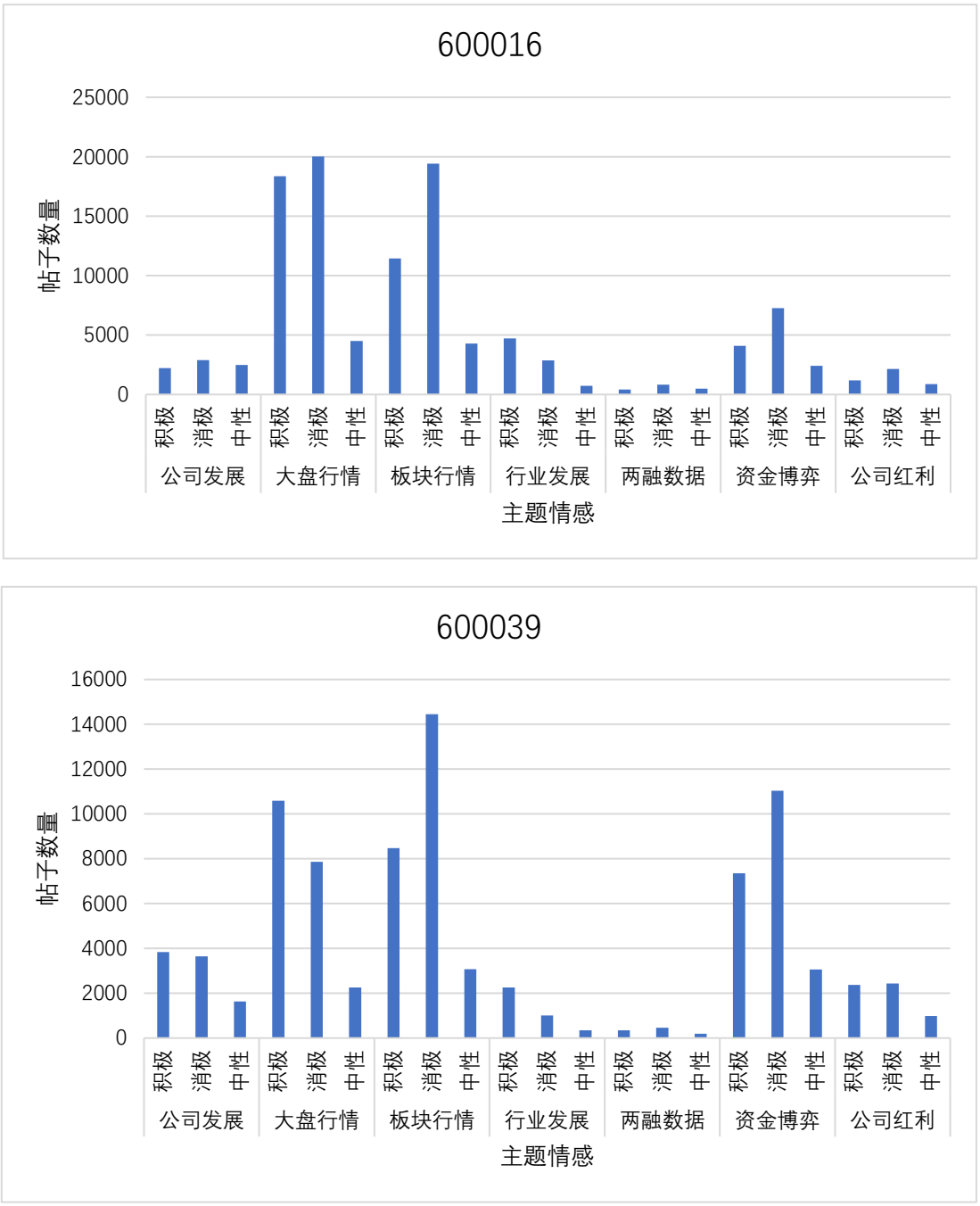

图 11 部分主题情感分析结果

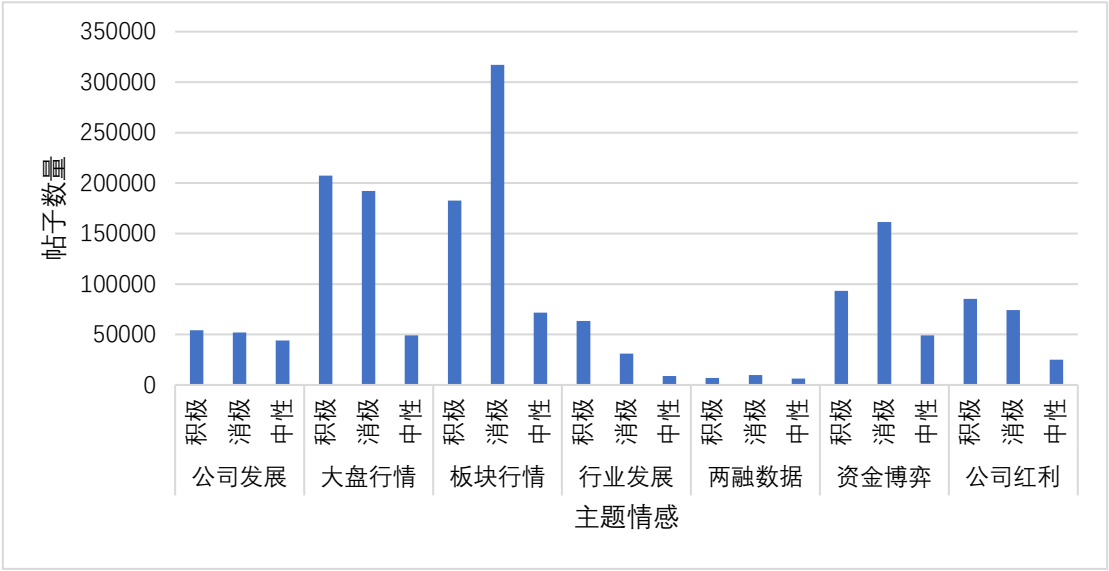

图 12 所有主题情感分析结果

4.3 建立回归模型

4.3.1 变量定义

(1) 超额收益（$Aret_{i,t,t+20}$和$Aret_{i,t,t+60}$）。

由于选取的样本均为上证 50 指数成分股，本文选取上证 50 指数作为基准指数，日超额收益定义为个股每天的收益率减去上证 50 指数收益率，将日超额收益率累加，即可得到股票一个月累计超额收益率$Aret_{i,t,t+20}$和三个月累积超额收益率$Aret_{i,t,t+60}$，将其作为本文构建模型的主要被解释变量。

(2) 情绪指标（Emo_1和Emo_2）。

根据上文得到的情感分析结果，参考金德环等人[45]的做法，本文用前 20 天各主题内的帖子情绪构建如下情绪指标：

$$Emo_{i1} = \log\frac{1+Pos_i}{1+Neg_i} \tag{1}$$

$$Emo_{i2} = \frac{Pos_i - Neg_i}{Neg_i + Pos_i + Neu_i} \tag{2}$$

其中，Pos_i, Neg_i, Neu_i分别对应前 20 天中主题i的积极、消极和中性情绪的帖子数量。本文将Emo_1和Emo_2作为主要解释变量，Emo_1和Emo_2越大，表明投资者越乐观。

(3) 控制变量

除构建的情绪指标外，参考 Chen 等人[6]的做法，本文在回归模型中添加了市值、账面市值比、个股月报告数、个股 10 日相对报告数、一致预期评级强度和一致预期净资产收益率等控制变量。变量定义如表 4 所示

表 4 变量定义及说明 [2]

变量	符号	说明
被解释变量	$Aret_{i,t,t+20}$	个股一个月累积超额收益率
	$Aret_{i,t,t+60}$	个股三个月累积超额收益率
解释变量	Emo_1	见（1）式
	Emo_2	见（2）式
控制变量	$Volatility$	波动率：为过去一个月收益率的标准差
	$Size$	市值：取当日收盘市值的对数
	BM	账面市值比：账面价值为公司定期报告披露的数据
	$Volume$	成交量：取当日成交量的对数
	$Report\_Num\_M$	个股月报告数：个股 30 日内卖方报告数量(不含预测表)
	$Relative\_Report\_Num\_10$	个股 10 日相对报告数：ln (1 + 10 日个股报告量/10 日 A 股报告总量) * 1000
	$Con\_Rating\_Strength$	一致预期评级强度：数值越大，买入信号越强
	Con_Roe	一致预期净资产收益率：100 * (个股一致预期净利润 / 个股一致预期净资产)

[2] 数据来源为 CCER 经济金融数据库、聚源数据字典和朝阳永续盈利预测数据库

4.3.2 模型构建

为了探究投资者情绪指标能够影响股票超额收益，本文构建如下回归模型：

$$Aret_{i,t,t+20}\left(Aret_{i,t,t+60}\right)=\alpha_{1,j}Emo_{i1,j,t}+\alpha_{2,j}Emo_{i2,j,t}+\beta_k ControlVariables_{ik,t}+c+\varepsilon_{i,t}(3)$$

$$Aret_{i,t,t+20}\left(Aret_{i,t,t+60}\right)=\alpha_1 Emo_{i1,t}+\alpha_2 Emo_{i2,t}+\beta_j ControlVariables_{ij,t}+c+\varepsilon_{i,t}(4)$$

其中，分别选取个股一个月累积超额收益率$Aret_{i,t,t+20}$和个股三个月累积超额收益率$Aret_{i,t,t+60}$作为被解释变量，$ControlVariables_{j,t}$为一系列控制变量。(3)式为使用经 LDA 主题分类的情绪指标构建的回归模型，主题编号由下标j表示，(4)式为直接使用未经主题分类的情绪指标构建的回归模型。

4.3.3 回归结果分析

回归结果如表 5 所示。第（1）列和第（2）列的被解释变量分别为一个月和三个月股票超额收益率。在添加了一系列影响股票收益率的控制变量后，第（1）列中大盘行情Emo_1和Emo_2的回归系数分别为 17.729 和 40.687，在 5%的统计水平下显著；板块行情Emo_1和Emo_2的回归系数分别为 3.7431 和 10.085，在 10%的统计水平下显著；行业发展Emo_1和Emo_2的回归系数分别为 1.4761 和 3.8609，在 5%的统计水平下显著；公司红利Emo_1和Emo_2的回归系数分别为 1.8574 和 4.837，在 1%的统计水平下显著。可见，对股票短期收益率影响较大的主题为“大盘行情”、“板块行情”、“行业发展”和“公司红利”。这与直观经验是相符的，短期内，散户投资者讨论的诸如上涨、下跌等行情走势较容易形成共识，对市场影响较大，造成股价短期波动。“行业发展”主题下多是讨论最新的行业动态，当一个行业利好出现时，同样会较快反映到行业内公司股价上。由于公司分红派息时常常会导致股价异动，因此公司红利主题的回归系数最显著，与实际金融市场的反应一致。第（2）列中公司发展Emo_1和Emo_2的回归系数分别为 12.918 和 44.372，在 1%的统计水平下显著；公司红利Emo_1和Emo_2的回归系数分别为 6.961 和 21.184，在 5%的统计水平下显著。相较第（1）列的回归系数，“公司发展”和“公司红利”主题系数显著增大，而“大盘行情”主题系数减小。可见，当考虑股票长期收益率时，行情类主题的影响变小，真正关注公司未来基本面的“公

司发展”主题的影响变大，而“公司红利”主题的影响依然显著。考虑未经主题分类的情绪指标回归系数，第（3）列Emo_1和Emo_2的回归系数为 10.904 和 27.856，在 10%的统计水平下显著，第（4）列Emo_1和Emo_2的回归系数为 80.551 和 218，分别在 5%和 1%的统计水平下显著。可见，未经主题分类的情绪指标同样对长期股票超额收益率的影响更大，但相比经主题分类的情绪指标回归结果，系数显著性降低，模型拟合效果（R^2）变差。综合来看，经主题分类的投资者情绪能够提取股票文本中的价值信息，各主题情绪的回归系数均为正数，表明股票超额收益与投资者情绪指标正相关，对于不同时间长度的超额收益率的影响因素也在变化，行情类信息影响短期收益，公司发展类信息影响长期收益，而公司红利信息对二者都有影响，与预期相符。

表 5 超额收益与情绪指标回归结果

变量	（1） $\mathrm{Aret}_{i,t,t+20}$	（2） $\mathrm{Aret}_{i,t,t+60}$	（3） $\mathrm{Aret}_{i,t,t+20}$	（4） $\mathrm{Aret}_{i,t,t+60}$
Emo_1			10.904*	80.551**
			(6.2894)	(32.794)
Emo_2			27.856*	218***
			(15.132)	(78.899)
公司发展Emo_1	0.3723	12.918***		
	(0.7914)	(4.001)		
公司发展Emo_2	0.9314	44.372***		
	(2.2052)	(11.149)		
大盘行情Emo_1	17.729**	10.441		
	(7.5828)	(38.338)		
大盘行情Emo_2	40.687**	23.78		
	(17.237)	(87.146)		
板块行情Emo_1	3.7431*	12.039		
	(0.7914)	(11.02)		
板块行情Emo_2	10.085*	34.557		
	(2.1797)	(28.102)		
行业发展Emo_1	1.4761**	2.6018		
	(0.657)	(3.3219)		
行业发展Emo_2	3.8609**	12.003		
	(1.6711)	(8.449)		

两融数据Emo_1	0.6931*	0.4074		
	(0.3689)	(1.865)		
两融数据Emo_2	0.0788	10.195*		
	(1.0747)	(5.4335)		
资金博弈Emo_1	1.9483	12.873		
	(1.6727)	(8.4569)		
资金博弈Emo_2	6.177	38.339*		
	(4.3282)	(21.883)		
公司红利Emo_1	1.8574***	6.961**		
	(0.6214)	(3.1418)		
公司红利Emo_2	4.837***	21.184**		
	(1.6622)	(8.4038)		
Volitility	-41.677***	-171.78***	-23.72***	-105.29**
	(8.6589)	(43.779)	(8.0408)	(41.926)
Size	0.4471	-12.987***	0.4047	-15.532***
	(0.6472)	(3.2722)	(0.615)	(3.2067)
BM	-0.000009078	-0.0009	-0.0000329	-0.0016***
	(0.0000724)	(0.7914)	(0.00006468)	(0.0003)
Volume	0.0322	0.3125	-0.0327	0.06
	(0.0798)	(0.7914)	(0.0817)	(0.4258)
Report_Num_M	0.0107	-0.39***	0.0153	-0.3487***
	(0.0185)	(0.7914)	(0.0192)	(0.1)
Relative_Report_Nu	0.0721	0.9286***	0.0643	0.9738***
	(0.0564)	(0.7914)	(0.059)	(0.3076)
Con_Rating_Strengtl	-0.9879	-2.641	-1.7979**	-6.3615*
	(0.7096)	(0.7914)	(0.6982)	(3.6406)
Con_Roe	-0.067	0.3131	-0.0445	0.2943
	(0.0599)	(0.7914)	(0.057)	(0.2972)
Const	-6.1579	226.5***	-6.0054	274.93***
	(11.342)	(0.7914)	(10.62)	(55.372)
R^2	0.1990	0.3562	0.0704	0.2052

注：*、**和***分别表示 10%、5%和 1%的显著性水平

4.4 进一步研究

4.4.1 非对称效应

Baumeister(2001)[46]提出，在日常生活中，坏的事情比好的事情更有力量，与好的印象相比，坏的印象和坏的刻板印象更容易形成。在金融领域，相关学者[47]研究发现在针对特定公司的新闻报道中，消极词汇的比例能够预测公司股价的低收益，对于那些关注公司基本面的报道而言，相比正面消息，负面消息对收益和回报的可预测性更大。因此，本文进一步考察投资者情绪中的正面和负面情绪对股价超额收益的影响是否存在非对称效应，利用过去一个月帖子情绪构建如下正负情绪指标：

$$PosEmo_{j,t,t+20} = \frac{Pos_j}{Neg_j + Pos_j + Neu_j} \tag{5}$$

$$NegEmo_{j,t,t+20} = \frac{Neg_j}{Neg_j + Pos_j + Neu_j} \tag{6}$$

其中，下标j对应不同主题，$PosEmo_{j,t,t+20}$越大表示过去一个月主题j的正面情绪越多，$NegEmo_{j,t,t+20}$越大表示过去一个月主题j的负面情绪越多。构建相应回归模型如下：

$$Aret_{i,t,t+20}\left(Aret_{i,t,t+60}\right) = \alpha_{1j}PosEmo_{ij,t,t+20} + \alpha_{2j}NegEmo_{ij,t,t+20} + \beta_j ControlVariables_{ij,t} + c + \varepsilon_{i,t} \tag{7}$$

回归结果如表 6 所示。第（1）列中“两融数据”主题Neg回归系数为-2.4841，在 1%的统计水平下显著，而其Pos回归系数不显著，说明“两融数据”主题情绪存在非对称效应，其他主题情绪回归系数均不显著。第（2）列“公司发展”主题Pos和Neg回归系数分别为 19.041 和-35.552，较第（1）列系数显著增大，且在 1%的统计水平下显著，“大盘行情”和“公司红利”主题回归系数较第（1）列均显著增大，且显著性水平提升，说明正、负面情绪对长期股价的影响更显著。分主题看，投资者对公司发展持悲观态度会导致长期股价显著下跌，对大盘行情的态度也会影响公司长期股价，与常识相符。从“公司发展”、“大盘行情”和“公司红利”主题Pos和Neg系数来看，Pos系数均为正，Neg系数均为负，说明股票超额收益同时受正、负面情绪影响，且Neg系数绝对值均大于Pos系数绝对值，

证实了非对称效应的存在。综合来看，无论是影响的程度，还是影响的持续性，负面情绪对股票超额收益率的影响都更为显著。

表 6 非对称效应

变量	（1） $Aret_{i,t,t+20}$	（2） $Aret_{i,t,t+60}$
公司发展Pos	-0.6331 (0.8564)	19.041*** (3.7641)
公司发展Neg	-0.2581 (0.7604)	-35.552*** (3.3418)
大盘行情Pos	2.7318 (2.6497)	29.821** (11.645)
大盘行情Neg	1.7248 (2.7579)	-35.57*** (12.121)
板块行情Pos	-2.3649 (2.2565)	5.0726 (9.9173)
板块行情Neg	-3.0815 (2.1394)	1.217 (9.4027)
行业发展Pos	1.4987 (1.0204)	1.8736 (4.4849)
行业发展Neg	1.107 (1.1495)	-8.6831 (5.0524)
两融数据Pos	0.6972 (0.8652)	0.5112 (3.8028)
两融数据Neg	-2.4841*** (0.685)	-15.456*** (3.0106)
资金博弈Pos	-1.7 (1.8354)	8.341 (8.0666)
资金博弈Neg	-2.5929 (1.6513)	2.6139 (7.2577)
公司红利Pos	-1.3414 (0.9444)	8.3378** (4.1508)
公司红利Neg	-1.2177 (0.8707)	-11.87*** (3.8269)
Volitility	-40.94*** (8.9163)	-182.5*** (39.188)

Size	0.1251	-8.7629***
	(0.6848)	(3.0096)
BM	-0.000005812	-0.0005
	(0.00007664)	(0.0003)
Volume	0.0433	0.6394*
	(0.0834)	(0.3666)
Report_Num_M	0.0152	-0.2406***
	(0.0191)	(0.0838)
Relative_Report_Num_10	0.0438	0.839***
	(0.0576)	(0.253)
Con_Rating_Strength	-1.0697	-1.0837
	(0.7193)	(3.1612)
Con_Roe	-0.0184	0.5331
	(0.0635)	(0.279)
Const	1.2652	144.53***
	(11.997)	(52.726)
R^2	0.1569	0.4879

注: *、**和***分别表示 10%、5%和 1%的显著性水平

4.4.2 超预期盈余

前文已经证实经主题分类的投资者情绪可以预测股票超额收益，说明投资者情绪中包含一定的价值信息。然而，前文结论可能是由于在股吧论坛上投资者情绪的集体偏差，这种“羊群效应”导致股价与投资者情绪表现出一定的相关性。为了探究投资者情绪是否包含真正的价值信息，本文进一步考虑投资者情绪和超预期盈余的相关性，如果投资者情绪包含真正的价值信息，那么其对超预期盈余应具有一定的解释力。参考 Chen 等(2004)的做法，本文构建如下超预期盈余指标：

$$Exc_{i,t,t+20} = \frac{EPS - EPS_{i,t,t+20}}{Price} \tag{8}$$

$$Exc_{i,t,t+60} = \frac{EPS - EPS_{i,t,t+60}}{Price} \tag{9}$$

其中，EPS为公司定期报告披露的真实EPS，$EPS_{i,t,t+20}(EPS_{i,t,t+60})$为过去一

（三）个月卖方分析师的平均滚动一致预期EPS[3]。因此，$Exc_{i,t,t+20}(Exc_{i,t,t+60})$为过去一（三）个月的平均超预期盈余。构建如下回归模型：

$$Exc_{i,t,t+20}(Exc_{i,t,t+60}) = \alpha_{1,j}Emo_{i1,j,t} + \alpha_{2,j}Emo_{i2,j,t} + \beta_j ControlVariables_{ij,t} + c + \varepsilon_{i,t} (10)$$

回归结果如表 7 所示。第（1）列中“公司红利”主题Emo_1和Emo_2的回归系数为 0.1747 和 0.5513，在 5%的统计水平下显著大于零，说明投资者对公司未来发放红利的态度越乐观，公司获得超预期盈余的程度就越大。第（2）列中“公司发展”主题Emo_1和Emo_2的回归系数为 0.2448 和 0.5169，分别在 5%和 10%的统计水平下显著大于零，“行业发展”主题Emo_1和Emo_2的回归系数为 0.2802 和 0.6536，在 1%的统计水平下显著大于零，“公司红利”主题Emo_1和Emo_2的回归系数为 0.2534 和 0.651，在 1%的统计水平下显著大于零，且系数较第（1）列显著增大。这意味着从长期看，投资者越看好公司及其行业发展前景，对公司发放红利持乐观态度，公司盈利超预期程度越大。这与经济学常识相符，公司和行业发展前景是影响公司盈利能力的重要因素，对这些因素越看好，表明公司未来盈利水平越高，就会带来更多的分红派息。综合来看，投资者情绪中包含与公司超预期盈余相关的价值信息，且对公司长期盈利能力影响更大。

表 7 超预期盈余

变量	（1） $\mathrm{Exc}_{i,t,t+20}$	（2） $\mathrm{Exc}_{i,t,t+60}$
公司发展Emo_1	0.0018 (0.1048)	0.2448** (0.107)
公司发展Emo_2	-0.1811 (0.292)	0.5169* (0.2981)
大盘行情Emo_1	-0.3385 (1.0041)	-1.4446 (1.0251)
大盘行情Emo_2	1.2984 (2.2824)	3.9152* (2.3301)
板块行情Emo_1	-0.0517 (0.2886)	0.5025* (0.2947)
板块行情Emo_2	-0.1934 (0.736)	1.0734 (0.7514)

[3] 数据来源为朝阳永续盈利预测数据库

行业发展Emo_1	0.1638 (0.087)	0.2802*** (0.0888)
行业发展Emo_2	0.2584 (0.2213)	0.6536*** (0.2259)
两融数据Emo_1	0.0521 (0.0488)	-0.0508 (0.0499)
两融数据Emo_2	-0.1254 (0.1423)	0.2843* (0.1453)
资金博弈Emo_1	-0.3269 (0.2215)	0.1827 (0.2261)
资金博弈Emo_2	0.9277 (0.5731)	-0.2471 (0.5851)
公司红利Emo_1	0.1747** (0.0823)	0.2534*** (0.084)
公司红利Emo_2	0.5513** (0.2201)	0.651*** (0.2247)
Volitility	6.9385*** (1.1466)	0.3259 (1.1705)
Size	0.0085 (0.0857)	0.126 (0.0875)
BM	-0.000005117 (0.000009588)	-0.00001435 (0.000009788)
Volume	0.0027 (0.0106)	0.0092 (0.0108)
Report_Num_M	0.005** (0.0025)	0.0072*** (0.0025)
Relative_Report_Num_10	0.0113 (0.0075)	0.0042 (0.0076)
Con_Rating_Strength	0.4187** (0.094)	0.2583*** (0.0959)
Con_Roe	0.0406** (0.0079)	0.0463*** (0.0081)
Const	-0.6981 (1.5019)	-2.3019 (1.5332)
R^2	0.4259	0.4205

注: *、**和***分别表示 10%、5%和 1%的显著性水平

5 总结与展望

5.1 总结

随着各类社交平台和投资论坛的发展，越来越多的机构投资者开始将舆情分析应用到投资模型中，在笔者实习的量化私募中，为了应对传统因子的拥挤度问题，也开始在新闻舆情上挖掘新的因子，并取得了不错的效果。由于文本数据为非结构化数据，在此类数据的预处理上就拥有了很多的选择和工具，如何更快速、准确地挖掘文本信息成为近年的研究热点。

本文首先介绍了投资者情绪领域的相关理论和研究方法，再介绍文本挖掘领域的发展简史和主流模型，分析各类模型的优点与不足，采用全概率生成模型 LDA 作为本文的主要模型。其次，详细介绍了 LDA 主题模型的理论基础与推理过程，包括必要的概率论知识和经典的机器学习算法。最后，利用东方财富股吧论坛文本数据进行实证分析。本文提供了一种创新性的文本挖掘方法，将 LDA 主题模型应用至股吧论坛情感倾向分析上，试图探究与股票超额收益率相关性最高的主题情感类型，剥离出海量帖子背后真正导致股价变动的情绪。研究结论表明，（1）使用经主题分类的情绪指标能提取出海量股票文本中的价值信息，与股票超额收益率正相关；（2）不同主题情绪对股票超额收益影响的程度和持续性不同，行情类信息影响短期收益，公司发展类信息影响长期收益，而公司红利信息对二者都有影响；（3）主题情绪对股票超额收益率的影响存在非对称效应，“公司发展”、“大盘行情”和“公司红利”主题非对称效应最显著；（4）主题情绪能有效解释公司超预期盈余，投资者越看好公司及其行业发展前景，对公司发放红利持乐观态度，公司盈利超预期程度越大。

5.2 展望

虽然本文提出的文本主题分类情感分析方法针对东方财富股吧文本数据取得了一些效果，但该方法仍存在不少问题：

(1) 没有预先构建针对股票信息的主题词典。股吧帖子涉及的话题数量非常广，投资者发表的帖子有的逻辑严谨，有的天马行空，若要实现准确的

文本主题分类，还需尽可能利用文本挖掘或人工筛选方法预先构建主题词典，以便更具针对性的进行主题模型训练。

(2) 本文没有在情感倾向分析上做太多创新。应当注意到，即使现有的开源工具能够在文本情感分析上有较好的表现，针对股市文本的分析也还存在很多偏差。由于股市文本包含很多金融领域的术语，普适性的情感分析可能无法识别这类特殊情况，因此，应事先构建针对股票市场的情感词典，将其与开源工具结合，提高情感倾向分析的准确性。

(3) 数据样本存在局限性。本文选取的数据样本不够庞大，并且选取的样本都是上证 50 指数成分股，没有对小盘股进行分析，后续可以将小盘股也纳入研究对象，论证模型的普适性。

参考文献

[1] Fama E F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work[J]. The Journal of Finance, 25(2): 383-417.

[2] Shiller R J, Fischer S, Friedman B M. 1984. Stock prices and social dynamics[J]. Brookings papers on economic activity, 1984(2): 457-510.

[3] Hung M, Li X, Wang S. 2015. Post-earnings-announcement drift in global markets: Evidence from an information shock[J]. The Review of Financial Studies, 28(4): 1242-1283.

[4] 刘永泽, 高嵩. 2015. 证券分析师行业专长, 预测准确性与市场反应[J]. 经济管理, 6: 87-97.

[5] Bagnoli M, Beneish M D, Watts S G. 1999. Whisper forecasts of quarterly earnings per share[J]. Journal of Accounting and Economics, 28(1): 27-50.

[6] Chen H, De P, Hu Y J, et al. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media[J]. The Review of Financial Studies, 27(5): 1367-1403.

[7] 中国证券投资者保护基金有限责任公司. 2019 年度全国股票市场投资者状况调查报告 [R]. 2020.

[8] Baker M, Wurgler J. 2007. Investor sentiment in the stock market[J]. Journal of economic perspectives, 21(2): 129-152.

[9] 郑振龙, 林璟. 2015. 沪深 300 股指期货定价偏差与投资者情绪[J]. 数理统计与管理, 34(06): 1129-1140.

[10] Beer F, Zouaoui M. 2011. Measuring Investor Sentiment in the Stock Market[J]. Journal of Applied Business Research, 29.

[11] 鹿坪, 冷军. 2017. 投资者情绪与盈余管理——基于应计盈余管理与真实盈余管理的实证研究[J]. 财经问题研究, (02): 88-96.

[12] Lavrenko V, Schmill M, Lawrie D, et al. Language models for financial news recommendation [M]. Proceedings of the ninth international conference on Information and knowledge management. McLean, Virginia, USA; Association for Computing Machinery. 2000: 389–396.

[13] Fung G, Yu J, Lam W. News Sensitive Stock Trend Prediction [M]. 2002.

[14] Antweiler W, Frank M Z. 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards[J]. The Journal of Finance, 59(3): 1259-1294.

[15] Frisbee B. The Predictive Power of Financial Blogs, F 2010, 2010 [C].

[16] Wysocki P. Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards [M]. 1998.

[17] Tumarkin R, Whitelaw R. 2001. News or Noise? Internet Postings and Stock Prices[J]. Financial Analysts Journal - FINANC ANAL J, 57: 41-51.

[18] Das S, Martinez-Jerez F, Tufano P. 2005. eInformation: A Clinical Study of Investor Discussion and Sentiment[J]. Financial Management, 34.

[19] 李玉梅. 基于互联网评论的股票市场趋势预测 [D]; 哈尔滨工业大学, 2012.

[20] 宋敏晶. 基于情感分析的股票预测模型研究 [D]; 哈尔滨工业大学, 2013.

[21]莫倩, 张渝杰, 胡航丽, et al. 2011. 一种混合的股评观点倾向性分析方法[J]. 计算机工程与应用, 47(19): 222-225.
[22]杨伟杰, 马博渊, 刘雯. 2014. 基于意见目标句抽取的中文股评情感分析方法[J]. 计算机仿真, 31(03): 431-436.
[23]张对. 2015. 网络股评影响股市走势吗——基于股票情感分析的视角[J]. 现代经济信息, (01): 355-357.
[24]Baeza-Yates R A, Ribeiro-Neto B. Modern Information Retrieval [M]. Addison-Wesley Longman Publishing Co., Inc., 1999.
[25]Salton G, McGill M J. Introduction to modern information retrieval [M]. McGraw-Hill, 1983.
[26]Deerwester S, Dumais S T, Furnas G W, et al. 1990. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 41(6): 391-407.
[27]Papadimitriou C H, Tamaki H, Raghavan P, et al. Latent semantic indexing: a probabilistic analysis [M]. Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. Seattle, Washington, USA; Association for Computing Machinery. 1998: 159–168.
[28]Hofmann T. Probabilistic latent semantic indexing [M]. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. Berkeley, California, USA; Association for Computing Machinery. 1999: 50–57.
[29]Blei D M, Ng A Y, Jordan M I. 2003. Latent dirichlet allocation[J]. J Mach Learn Res, 3(null): 993–1022.
[30]Lu Y, Zhai C. Opinion integration through semi-supervised topic modeling [M]. Proceedings of the 17th international conference on World Wide Web. Beijing, China; Association for Computing Machinery. 2008: 121–130.
[31]Hong L, Davison B D. Empirical study of topic modeling in Twitter [M]. Proceedings of the First Workshop on Social Media Analytics. Washington D.C., District of Columbia; Association for Computing Machinery. 2010: 80–88.
[32]Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents [M]. Proceedings of the 20th conference on Uncertainty in artificial intelligence. Banff, Canada; AUAI Press. 2004: 487–494.
[33]Titov I, McDonald R. Modeling online reviews with multi-grain topic models [M]. Proceedings of the 17th international conference on World Wide Web. Beijing, China; Association for Computing Machinery. 2008: 111–120.
[34]Titov I, McDonald R T. A Joint Model of Text and Aspect Ratings for Sentiment Summarization [M]. Association for Computational Linguistics. 2008: 308-316.
[35]Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid [M]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts; Association for Computational Linguistics. 2010: 56–65.
[36]Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews [M]. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California; Association for Computational Linguistics. 2010: 804–812.

[37]Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis [M]. Proceedings of the fourth ACM international conference on Web search and data mining. Hong Kong, China; Association for Computing Machinery. 2011: 815–824.
[38]Lin C, He Y. Joint sentiment/topic model for sentiment analysis [M]. Proceedings of the 18th ACM conference on Information and knowledge management. Hong Kong, China; Association for Computing Machinery. 2009: 375–384.
[39]Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs [M]. Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada; Association for Computing Machinery. 2007: 171–180.
[40]孙艳, 周学广, 付伟. 2013. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报(自然科学版), 49(01): 102-108.
[41]Dickey J M. 1983. Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses[J]. Journal of the American Statistical Association, 78(383): 628-637.
[42]Nigam K, McCallum A K, Thrun S, et al. 2000. Text classification from labeled and unlabeled documents using EM[J]. Machine learning, 39(2-3): 103-134.
[43]Popescul A, Pennock D M, Lawrence S. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments [M]. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. Seattle, Washington; Morgan Kaufmann Publishers Inc. 2001: 437–444.
[44]Griffiths T L, Steyvers M. 2004. Finding scientific topics[J]. Proceedings of the National academy of Sciences, 101(suppl 1): 5228-5235.
[45]金德环, 李岩. 2017. 群体智慧:同伴观点与价值发现——来自社交媒体的经验证据[J]. 经济管理, (12): 157-173.
[46]Baumeister R F, Bratslavsky E, Finkenauer C, et al. 2001. Bad is stronger than good[J]. Review of general psychology, 5(4): 323-370.
[47]Tetlock P C, Saar-Tsechansky M, Macskassy S. 2008. More than words: Quantifying language to measure firms' fundamentals[J]. The Journal of Finance, 63(3): 1437-1467.

致谢

本文是在我的导师张科老师的耐心指导下完成的。在论文撰写的全过程中，是导师适时的点拨和指导，帮助我快速理解这一领域的相关知识，今后我也将以导师为学习榜样，严谨治学、持之以恒，在研究生阶段继续自己的科研学习。

我曾无比向往的四年本科生活就要走到尾声了，这四年里有太多帮助过、陪伴过我的人。感谢本科阶段的各位老师们，带领我走进金融学的殿堂，让我找到自己的兴趣所在。感谢辅导员老师，在我困惑迷茫时为我指明前进的方向，在我升学道路上给予有力的支持。感谢我的女朋友一直以来的关心和陪伴，谢谢你让我学会如何珍惜、如何去爱。感谢朝夕相处的同学们，特别是我的室友们，陪我度过了难忘的四年本科生活。

在此特别要感谢我的父母，在我失落时给我的陪伴与鼓励，在我陷入自我怀疑时第一时间为我排忧解难，是你们成就了今天的我。

最后预先向负责审阅论文和各位答辩评审老师表示衷心的感谢！