分类号_____ 密级_____

U D C_____ 编号_____

厦 门 大 学

博 士 后 研 究 工 作 报 告

社交媒体大数据、投资者情绪与股票收益预测

王建新

工作完成日期：2016 年 6 月

报告提交日期：2016 年 6 月

厦门大学

2016 年 6 月

社交媒体大数据、投资者情绪与股市收益预测

Social Media, Big Data, Investor Sentiment and Stock Return Prediction

博　士　后　姓　名　王建新

流动站（一级学科）名称　工商管理

专　业（二级学科）名称　财务学

研究工作起始时间　2013 年　8 月

研究工作期满时间　2016 年　6 月

厦 门 大 学

2016 年 6 月

厦门大学博士后研究工作报告著作权使用声明

本人完全了解厦门大学有关保留、使用博士后研究工作报告的规定。厦门大学有权保留并向国家主管部门或其指定机构送交该报告的纸质版和电子版，有权将该报告用于非赢利目的的少量复制并允许该报告进入学校图书馆被查阅，有权将该报告的内容编入有关数据库进行检索，有权将博士后研究工作报告的标题和摘要汇编出版。保密的博士后研究工作报告在解密后适用本规定。

本研究报告属于：　　1、保密（　），　2、不保密（　）

纸本在　　年解密后适用本授权书；

电子版在　　年解密后适用本授权书。

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘要

近些年来，Web2.0 技术臻于成熟。随着互联网技术的飞速发展和移动电子设备的不断普及，各类形式的社交媒体不断涌现出来，并得到了广泛的应用。庞大的社交媒体用户群体每天在各类平台上表达自己的观点、见解、心情，这些信息都被自动记录下来，被称作用户生成数据（User Generated Data）。社交媒体上天然的海量数据为学术研究提供了良好的数据来源。社交媒体大数据被尝试应用于各个领域。本文从中文社交媒体大数据出发，以投资者情绪为桥梁，系统性分析社交媒体大数据对金融市场的预测作用以及大数据本身的偏差性问题。本文在对已有的相关研究和理论进行综述后，对我国社交媒体的发展现状进行了分析，在此基础上重点进行了如下三个方面的实证研究。

第一，检验了社交媒体情绪对个股收益率、交易量和波动率的预测作用。以股票简称为关键词抓取了腾讯微博平台上讨论中证 100 指数成分股的相关文本数据，通过股票简称界定个股投资者和潜在投资者发表的文本信息，并对文本信息进行情绪分析，从而获得了个股微博情绪。利用个股微博情绪对个股的预测作用进行了检验，发现个股微博情绪对个股日异常收益率不存在显著的预测作用；个股超额收益率对未来微博投资者情绪存在显著的预测作用；个股微博情绪对个股交易量存在显著的预测作用，对个股波动率则不存在预测作用。另外，微博发布量对个股异常收益率和个股交易量都存在显著的预测作用。

第二，检验了社交媒体情绪对股票市场指数收益率的预测作用。分别抓取了腾讯微博平台和雪球投资论坛上的与股票市场相关的文本数据，在对文本的情绪水平进行分析之后，对比分析了两个平台的社交媒体情绪对指数收益率的预测作用。实证发现：从腾讯微博上获得去投资者情绪能够在短期预测指数收益，而从雪球微博上获取的情绪则不存在预测作用。由于腾讯微博是大众化的微博平台，而雪球微博是专门的投资者交流平台，我们认为这种预测的差异性可能是由于社交媒体平台上存在市场操纵者导致的。

第三，针对前文研究结论中社交媒体整体情绪所表现出的弱预测性，对社

交媒体数据的偏差性进行了检验。通过收集“雪球”股票论坛中的讨论数据构建投资者情绪指标，实证研究发现：（1）无论是在日度、周度还是月度频率上，整体的投资者情绪对股票市场收益均不存在预测作用，表明社交媒体股票大数据存在偏差；（2）社交媒体中存在市场操纵者，“市场操纵”用户的情绪能够引领“非市场操纵”用户的情绪；（3）市场操纵者通过数据操纵达到了获利目的，即具有 “市场操纵”特征用户的情绪与未来股票市场收益负相关。这些研究结果表明：具有一定意图的投资者可能在社交媒体中发布具有倾向性的甚至虚假的信息，从而导致整体数据产生系统性偏差。

本文的研究具有一定的探索性，研究的主要创新和贡献有如下几个方面：

第一，对投资者情绪对个股收益的预测作用进行探索。以往对于投资者情绪进行的研究大多是对整个股票市场的预测，或者是将投资者情绪作为调节变量考察其对不同特征股票的影响的敏感性。主要原因在于单只股票的投资者的情绪很难测量。本研究借助于微博这一新兴的社交网络，利用投资者在微博上的发言内容区分每只股票的投资者和潜在投资者，并根据微博内容对每只股票的投资者的情绪进行分析，从而实现了投资者情绪对个股收益的预测作用的研究。

第二，丰富了新兴的网络社交媒体在金融市场中的应用研究，特别是对基于中文文本的社交媒体大数据对金融市场的预测性进行了系统性检验。随着web2.0 技术的不断发展，社交网络被大众所广泛应用，这些社交媒体记录下了大量的个人数据，为科学研究提供了良好的原始数据。本研究从微博角度分析股票市场上的投资者的情绪，并利用微博投资者情绪对个股收益率和市场收益率的预测作用进行了分析，这对新兴的社交网络如何应用于金融市场起到一定的贡献和补充作用。

第三，本文对社交媒体股票大数据的偏差性进行了检验。大数据得到广泛应用的一大好处是其在较大的程度上克服了抽样误差，这常常是得益于大数据的全数据性。然而，由于大数据本身并非是基于科学研究的目的而进行设计的，因此大数据在生成过程中就可能存在系统性的偏差。本文利用社交媒体股票大

数据，借助于投资者情绪对股票市场收益的理论预测性，对社交媒体大数据在生成过程中的偏差存在性进行了检验。研究结论对大数据在应用过程中的审慎性要求提供理论依据。

关键词： 社交媒体、大数据、投资者情绪、股票收益预测

Abstract

Benefit both from the rapidly developing of the technology of Web2.0 and the wide spread use of mobile device, many kinds of social media platforms, like Twitter and Facebook, are emerging during the past couples of years. Those social media platforms attracted billions of users worldwide. Huge amount of users are posting their views, opinions, emotions etc. on social media every single day. All of the posted messages are recorded on the platforms automatically, which is called User Generated Data. Big data on social media platform, seen as natural research data, is trying to be applied to every field in the academia. Based on the Chinses text data on social media and investor sentiment, this paper systematically analyses the prediction effect of social media data on stock market, and also examine the sample bias of big data on social media. After reviewing the related literature and describing the development of social media in China, we conduct the following empirical analysis.

Firstly, this paper examine whether social media sentiment could predict individual stock return, trading volume and volatility. We obtain stock related messages on “Tencent” microblog platform. We then identify stock-specific messages by stock name and analyze the text content of those messages to extract stock-specific investor sentiment. We find that stock-specific sentiment failed to predict the daily abnormal stock return, in contrast, abnormal stock return can predict stock-specific sentiment. Additionally, stock-specific sentiment can predict trading volume but not volatility. Besides, Microblog posting volume has significant prediction effect on both individual stock abnormal return and trading volume.

Secondly, we examine whether social media sentiment could predict stock marker return. We obtain investor posting messages from both “Tencent” and “Snowball” platform. On analyzing sentiment level from text messages, we examine the prediction power of social media sentiment on aggregate stock market return level. We find that “Tencent” sentiment can predict stock index return in the short period, while “Snowball” sentiment can not predict stock index return. We

believe that the weak predictability is caused by marker manipulators who also post in the social media.

Thirdly, we test whether there is sample bias in the social media big data. On measuring the investor sentiment in use of a large-scale(649636) discussion samples from the "Snowball" internet stock forum, we find: first of all, that the aggregate investor sentiment fails to predict the stock market return, in daily, weekly or monthly level. Secondly, market manipulators try to manipulate the data generating process on social media platform, the Granger Causal Test shows that manipulator's sentiment is the Granger reason of the non-manipulator's sentiment. Finally, manipulator's sentiment negatively predicts the market return in short term, which indicates that the manipulators do gain abnormal return from the manipulating behavior. The empirical results indicate that the "pump and dumpers" are trying to manipulate the market price thorough posting purposely misleading information, which will bias the aggregate data systematically.

Possible innovation and contribution of this research are as follows:

1. We study the prediction effect of investor sentiment on stock return on the individual stock level. Since stock-specific investor sentiment can not be measured easily, most of the past researches on investor sentiment focus on the aggregate sentiment and market level. Taking the advantage of social media, we innovatively measure stock-specific investor sentiment by identifying investor and potential investor's messages with stock name mentioned in the messages. Thus we can analyze the relation between investor sentiment and stock return on the individual stock level.

2. This study enrich the literature on the relation between social media and stock market, especially, we examine this relation based on Chinese social media and Chinese text which is rare, at least according to our knowledge. The technology of Web2.0 and social media industry is growing very fast, the big and natural data on social media supply opportunity to the investor sentiment research. We obtain Chinese based text data from social network in China and study the issue above, which is complementary to this research field.

3. We innovatively examine the sample bias of the big data on social media. Big

data, often seen as full sample, overcomes the sampling error to some extent. Nonetheless, one of the concerns of big data is they are not produced based on the purpose of scientific research, thus, there could be bias during the data generating process. This paper make in use of the social media data related to stock market, based on the theoretically prediction of investor sentiment to stock return, examine the sample bias during the data generating process. Big data researches should be aware of the type of bias formed in the data-generating process.

Keywords: Social Media; Big Data; Investor Sentiment; Stock Return Prediction

目　录

1 绪论

1.1 研究背景

自从 Delong 等（1990）在 DSSW 噪声交易者模型中引入投资者情绪以来，20 多年的时间里，对于投资者情绪的研究层出穷。投资者情绪也因此成为行为金融学的一块重要基石。在 Delong 等（1990）从理论上证明投资者在决策中受到情绪的影响之后不久，Lee 等（1991）就为该理论找到了实证证据，利用从封闭式基金折价的视角进行了验证。20 余年间关于投资者情绪的主要发现有：投资者情绪能够系统性地造成股票价格偏离基础价值、投资者情绪对未来股票价格存在一定的预测力、投资者情绪对股票价格的预测能力在横截面上表现出差异，如具有小规模、年轻、高波动性特征的股票对投资者情绪更敏感。

对于投资者情绪的实证研究，投资者情绪如何测量是关键因素和挑战。已有的研究主要通过以下几种方式衡量投资者情绪指标。第一种方式主要从金融市场本身出发，采用与股票交易相关的变量来代理投资者情绪，如封闭式指数基金折价率、换手率、IPO 首日回报率、月度股票市场新增开户数等。在初期的研究中学者往往采用其中的一种或几种指标作为投资者情绪的代理变量，Baker 等（2006）首次采用主成分分析的方法将该类指标进行了综合和加权，采用一揽子金融市场变量的方式更好的衡量投资者情绪。第二种方式则采用与投资者情绪相关的间接的变量进行测量，也就是那些在宏观上能够影响群体情绪的因素，如天气的好坏、体育赛事的输赢、季节性情绪失调(SAD，seasonal affective disorder)等，用这些外生能够影响群体情绪的变量来代理整个投资者群体的情绪。第三种方式最为直接，利用问卷调查的方式获取投资者情绪水平，如利用 Investors' Intelligence(II)的调查数据来度量投资者情绪，也有使用消费者调查数据或对其他种类人群的调查数据作为投资者情绪的代理变量的。

以上三种方式，使用最为广泛的是第一种方式，第二种方式的好处是具有较强的外生性，第三种方式较为直接。但三种方式都或多或少的存在一些弊端，如采用金融市场中的指标进行衡量可能存在较强的内生性问题，通过调查的方式获取数据成本又十分高昂。更为重要的一点是，所有三种方式基本都只能对

整体的投资者情绪进行衡量，因此也只能局限于对整体金融市场的影响的分析上。

随着互联网的出现和不断发展，学者们开始考虑和尝试利用互联网数据更加有效和经济地衡量投资者情绪。早期的尝试主要来自于雅虎网络讨论板的文本数据（Yahoo! Finance）。Wysocki（1999）最早利用雅虎网络留言板上的信息对股市收益进行预测。Antweiler 和 Frank（2004），他们使用的雅虎金融论坛的数据，发现与股票相关的讨论信息能够预测股票市场的波动，对于股票收益的预测并不具有经济意义上的显著性。然而这些利用网络讨论文本作为数据源的早期文献并非直接诉诸于情绪理论来解释其实证结果，更多是从信念和噪声的角度。Das 和 Chen（2007）提出了一种从网络信息中提取情绪变量的文本分析方法，并且利用股票价格数据对其有效性进行评估，但是并没有将其用于预测股价。此外，Tetlock（2007）和 Tetlock 等（2008）也借助内容分析法分别研究了网络信息内容如何影响指数和个股的价格行为，但是他们使用的是财经媒体的新闻数据。

最近七、八年来，基于 Web2.0 的互联网技术和应用得到了飞速发展。在博客发展和兴盛了几年之后，以 Facebook、Twitter、Linkin 等为代表的社交网络的出现，彻底改变和颠覆了传统社交的认知。国内社交网络的发展势头也十分迅猛。以微博为例，2010 年，国内微博像雨后春笋般崛起，新浪、腾讯、搜狐和网易四大门户网站均开设微博。2012 年 1 月，据中国互联网络信息中心报告显示，截至 2011 年 12 月底，我国微博用户数达到 2.5 亿，网民使用率为 48.7%。据《第 31 次中国互联网络发展状况统计报告》显示，截至 2012 年 12 月底，我国微博用户规模为 3.09 亿，网民中的微博用户比例达到 54.7%。社交网络为投资者表达自己的看法和见解提供了可能。投资者每天都在微博等社交网络上表达着自己的行为和心理状态，这些大数据特征信息为研究者提供了天然而原始的情绪分析数据。

学术领域对这些丰富的数据资源给予了充分的重视。最近几年，利用这些新兴的社交媒体数据，特别是微博（Twitter）数据进行的研究层出不穷，其中很大一部分是对投资者情绪进行的研究。比较具有代表性的如，Bollen, Mao 和 Zeng 于 2011 年在《Journal of Computational Science》发表的关于推特数据用于股市预测的研究。通过内容分析技术，他们从 970 万条推特（Twitter）信息中

提取出“镇静”、“警惕”、“肯定”、“活力”、“友善”、“幸福”等这六种与股市无关的人类情感测度，发现仅有“镇静度” 指数能够提前预测道琼斯工业平均指数的走向，准确率高达 87.6%。这篇文献在学术界引发了运用大数据技术预测股票市场的研究兴趣，近年来涌现出大量的跟踪研究，该文自发表以来谷歌学术引用次数已经超过 1800 次。

1.2 研究动机

对于这些已有的利用社交媒体文本数据对投资者情绪进行的研究主要存在以下三个方面的问题，从而也是本文的研究动机所在。

第一，利用中文文本进行的研究较为缺乏。

对于微博情绪是否能够预测股票收益这一命题，国内研究尚处于起步阶段。程琬芸等（2013）以新浪微博作为数据源来构建投资者涨跌情绪指数，发现这一情绪指标无法帮助预测证券市场指数的收益。赖凯声等（2014）用 140 天的新浪微博数据构建了综合情绪指数，发现该指数是一阶平稳的序列，并且与上证指数之间存在着显著的协整关系。黄润鹏等（2015）也同样采用新浪微博生成情绪倾向的时间序列，发现在传统的支持向量机预测模型中加入情绪变量后，预测准确率从 54.56%提高到 68.18%，但是仍然只达到 10%的显著水平，这一结果基本与 Rechenthin 等（2013）相似。2015 年 7 月，一份欧洲央行的统计报告[①]声称，每日的推特（Twitter）情绪指数对预测美国、英国和加拿大的股市走势有明显作用，而对中国股市的预测能力则较为微弱。这一新闻被多家网络媒体转载，并且在网上引发热烈的讨论[②]。有人认为微博的情绪体现了群体的智慧，是大数据技术用于金融投资的有力证据。有人质疑该报告所选用的样本区间刚好是金融危机爆发后的三年，具有特定性，其研究结论不适合推广。也有人指出中国的语言和市场环境与英语国家差异太大，使得推特情绪对中国股市的预测力降低，甚至会出现相反的模式。与 Mao 等（2015）相比，国内的相关研究所采用的是本土化语言（非英语）所表达的微博情绪，同时也采用了不同的计

[①] Mao H, Counts S, Bollen J. Quantifying the effects of online bullishness on international financial markets[R] European Central Bank Statistics Paper Series No.9, 2015. https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp9.en.pdf?177000b829d4450b007f3d3a612cab18。

[②] http://wallstreetcn.com/node/221107。

量技术（而非普通回归），因此其结论可比性不高。本文试图采用最简单的线性预测模型，来验证 Bollen 等（2015）在英语国家的股票市场所得到预测结果是否可以在中国股市复制。

第二，已有的相关研究中，大多是对整个市场的投资者的情绪的分析，缺乏个股投资者情绪的分析。仍然以 Bollen（2001）为例，该研究利用 970 万条推特（Twitter）信息提取情绪，但这 970 万条微博信息是一般性的信息，也就是所有微博用户所发表的信息。因此，利用这些信息衡量的依然是整体社会人群的情绪，而非投资者的情绪。然而，社交媒体的一大好处就是可以对不同群体进行区分。本文尝试利用与股票相关的社交媒体信息提取情绪，这些信息在一定程度上可以被认为是投资者或潜在投资者所发表的信息，因此，首先可以将情绪具体到投资者群体。进一步地，我们认为通过对文本信息中所讨论的股票进行甄别，可以区分出每只股票的投资者和潜在投资者，从而可以构建个股投资者的微博情绪指标，并进行相关的预测。

第三，现有的研究大多没有考虑数据本身的偏差性，这也是迄今所有大数据研究所面临的一个问题和挑战。以社交媒体数据为例，这些海量数据确实可以被视为研究数据的良好来源。但同时需要注意的是，这些数据本身的生成并非基于研究目的，因此数据本身的随机性和偏差性需要检验。大数据的偏差性已经受到了学术界的重视。Lazer 等（2014）发表于《Science》的论文利用谷歌流感趋势表现出的预测偏差对大数据偏离事实的原因进行了分析，并指出：由于大部分正在被使用的大数据资源本身并不是按照科学研究的需要而设计和生产出来的，大数据本身的结构效度、信度以及独立性等基本问题不容忽视。除了由于数据检索、处理、数据匹配等可能造成的偏差外，数据生成过程中也可能出现的系统性偏差。该文还特别强调了社交媒体上的信息往往可能由于某些具有目的性的数据操纵而导致偏差，具有政治意图（如选举）的社交媒体数据操纵已经得到证实（Mustafaraj & Metaxas，2010；Ratkiewicz et al.，2011），而基于经济意图（如金融市场操纵）的社交媒体数据操纵行为是否存在还处于推测阶段，需要经验证据的支持。因此，本文对社交媒体中的股票大数据的偏差性进行了检验。