社交媒体对股票市场的影响研究
——基于文本挖掘的实证分析

摘要：股票市场的表现不仅反映了一国的经济发展状况和资本市场趋势，也影响着个人的投资决策、公司的经营策略，甚至国家的经济政策。本文使用社交媒体上的大数据对股票市场进行分析和预测，借助文本挖掘技术、机器学习算法及计量经济学方法探究社交媒体情绪对股票市场的影响。在对社交媒体上的本文进行处理后从多个维度构建社交媒体情绪指数，并与股票市场的表现进行实证研究。研究结果表明，社交媒体情绪和股票市场之间存在着较强的非线性相关关系，加入社交媒体情绪指数有助于对股票市场的解释和预测。

关键词：文本挖掘；情感分析；机器学习；社交媒体；股票市场

一、引言

股票市场是自由市场经济最重要的组成部分之一。股票市场为公司提供了良好的融资平台和有法律保障的融资环境，公司发行股票能够高效地筹措发展所需要的资金，实现规模化经营。投资者通过购买股票以及股票流转，分担公司精英风险的同时也享受着相应的权益，比如股票的增值以及公司的分红。

美国股票市场作为当今世界上最发达的股票市场，其基本涵盖了世界上全部的知名企业，涉及能源、材料、工业、农业、医药、消费品等数十个行业。丰富的股票交易种类、庞大的股票发行数量和首屈一指的市场规模使得美国股票市场表现不仅反映美国经济和政治的走向，也影响全球的政治经济发展。

股票市场是反映经济发展状况和资本市场趋势的晴雨表，股票市场的表现可以反映出资金供求状况、行业发展趋势、政治形势变化、公司经营状况等信息，对于宏观经济政策的调控以及公司自身的经营决策都有着重要意义。因此，对股票市场的分析和预测不仅可以帮助投资者避免损失，还有利于国家制定相关经济政策以及公司改善经营决策。

行为金融学理论运用社会学、心理学、行为学的研究成果研究投资者的非理性行为并从投资者的行为和心理来分析、解释资本市场的表现。行为金融学在数理金融学的基础上加以修正，将投资者的价值感受，即心理动机和个人行为纳入对金融市场波动的分析（Froot 和 Scharfstein，1990）。心理学研究表明情绪在人们决策过程中发挥着重要的作用，纳入了心理因素的行为金融学认为投资者非理性的心理和行为会影响股票市场的走势，并使股票价格偏离其本身的价值（Kahneman 和 Tversky，1997）。

作为人们记录和表达的载体，社交媒体中拥有大量的用户情绪和行为信息，为行为金融学分析提供了充足的数据来源。美国互联网的普及率高达 88.7%，数亿用户在社交媒体上记

录生活、交流观点，每天 Twitter 上有超过 5 亿条内容发布，超过 4000 万照片上传到 ins 上，Facebook 的全球用户超过 24 亿。与此同时，越多越多的研究注意到社交媒体的影响力，从社交媒体中挖掘信息并提炼投资者情绪逐渐成为金融学、计算机科学、管理学等领域的热门话题。

自新冠疫情爆发以来，多国股市出现大规模动荡。3 月 12 日美国三大股指开盘下跌超过 7%，集体触发熔断机制；巴西股市下跌 15%，一天内触发二档熔断机制；韩国股市下跌将近 11%.....目前至少 12 个国家主要股市触发熔断机制暂停交易，数十个股市进入“技术性熊市”。与此同时，Twitter、Facebook 上民众的担忧和焦虑情绪也愈演愈烈，因此我们有充分的理由假设社交媒体上的情绪和股票市场的表现具有相关性。

本文将投资者情绪拓展为公众情绪，以社交媒体上大数据为基础，借助网络爬虫、自然语言情感分析、机器学习算法等技术，探究社交媒体上的公众情绪与股票市场的关系。本文将先进的计算机科学与前沿的金融学理论结合，从而提高股票预测模型的精准度。精准地预测股票走势，不仅可以帮助投资者获得收益，还能够为相关经济政策的制定提供合理依据。

二、文献综述

股票市场的分析和预测一直是学术界和商业界关注的热点话题。早期的股票市场预测主要基于随机漫步理论和有效市场假说。随机漫步理论认为股票价格的变化类似于随机漫步的布朗运动，股票市场的波动具有随机性和无规律性，因此股票价格变动及其趋势是无法被预测的[4]。基于理性人假设的有效市场假说同样认为股价不可预测，因为在有效市场中所有的信息都已经及时、准确、充分地反映在股价走势中，投资者不能通过分析过往价格获得预期收益以上的超额收益。

对于股票市场的预测，经典金融学基于期望效用函数提出了诸多资本定价理论，包括 Markowitz（1952）的投资组合选择理论、Sharpe 的资本资产定价模型、以及 Stephen Ross（1990）的套利定价理论等等，但是这些建立在理性人假设基础上的股价预测模型并不能解释股票市场中的一些现象，比如股价长期反转效应（De Bondt，1985）和股权溢价之谜（Mehra 和 Prescott，1985），其预测的结果也常与现实背离。

行为金融学理论放松了理性经济人的假设，将行为人的心理因素考虑在内，并由此对原有的经典金融学理论加以修正。Baker（2007）的研究表明，投资者的情绪是可以被衡量的，且投资者的情绪波动对股票市场具有可辩别的重要影响。

DeLong, Shleifer, Summers 和 Waldmann（1990）提出的噪声交易模型（简称 DSSW 模型）是投资者情绪元素定价的先驱。噪声指与资产价值无关但影响其价格的信息（BLACK F，1986），即资产价格与价值之间的偏差，噪音交易指基于非理性的心理或与资产价值无关的信息进行的交易，噪声交易者即进行噪声交易的非理性投资者。DSSW 模型将噪声交易者引入资产定价模型并解释了噪声交易对资产价格的影响。

常用的投资者情绪度量方法主要包含两个步骤，第一步挑选市场整体换手率、封闭式基

金折价率、IPO 首日溢价率、新股发行数等市场指标作为情绪变量，第二步采用主成分分析法、偏最小二乘回归方法或动态因子建模方法构建综合的投资者情绪指数。之后再根据不同的研究目的选择合适的金融理论和统计模型，将投资者情绪指数作为变量加入其中，完成整个实证分析流程。

随着社交媒体的广泛应用，许多研究从社交媒体（Facebook、Twitter、新闻网站等）中提取相应指标来预测各种经济和商业指标的变化。例如，D.Gruhl（2005）使用博客、媒体和网站上的内容预测图书销售情况，G.Mishn（2006）对博客内容做情感分析从而预测电影票房，除此之外，谷歌搜索指数已经被证实可以预测流感趋势（Choi 和 Varian，2009）。这些研究表明，社交媒体中的内容具有一定程度的预测作用。

行为金融学理论认为投资者的情绪能够影响股票市场的表现，而社交媒体为广大投资者提供了交流的平台，因此许多研究从社交媒体中提取投资者情绪构建情绪指标。例如，Schumaker 和 Chen（2009）运用文本分析技术探究财经新闻和股票评论中的情感指数和股价变动的关系，Yigitcan Karabulut（2017）使用 Facebook 的国民幸福指数预测股价变动，Bollen J 和 Mao（2011）发现 Twitter 中的平静类情绪能在一定程度上预测道琼斯指数等。

相较于上述文献，本文主要有以下两点创新：

分别使用情感词典法和五种机器学习算法对文本进行情感分析，结果表明机器学习算法的准确率远高于情感词典法，其中朴素贝叶斯算法的表现最好，准确率高达 88%，为后续的研究提供了扎实的数据基础。

对比传统的线性回归模型和深度学习算法 LSTM，使用典型的非线性回归模型 LSTM（长短期记忆网络）构建的股价预测模型误差更小，表明社交媒体情绪和股票市场的表现之间存在着较强的非线性关系。

三、研究设计

本文综合运用金融学、计算机科学、计量经济学等多领域知识探究社交媒体对股票市场的影响。考虑到 Twitter 平台每月超过 3 亿的活跃用户数及其广泛影响力，选择 Twitter 作为数据来源构建社交媒体情绪指数能够较为准确地反映社交媒体用户的情绪。相较于道琼斯指数和纳斯达克指数，标普 500 的成分股涵盖范围更广、代表性更强、连续性更好，因此采用标普 500 指数能够较为精确地衡量美国股票市场表现。

理论分析方面，本文首先阐释了研究的背景和意义，并且通过文献调研对比了国内外相关论题的研究现状，从而制定合理的研究方法和技术路线，为接下来的实证研究打下理论基础。实证研究方面，本文选取在世界内具有广泛影响力的 Twitter 平台及标普 500 指数作为研究对象，借助文本挖掘技术、自然语言情感分析、机器学习算法及计量经济学方法实现数据的采集、处理以及模型的构建，从而探究社交媒体情绪对于股票市场的影响。

（1）网络爬虫技术

本文采用编程语言 Python 和网络爬虫技术获取 Twitter 上所有用户在一定时间内发布的

文本数据，作为后续用来分析的数据来源。

（2）自然语言处理

本文采取分词、词袋模型等方法将爬取到的非结构化文本内容进行清洗和分词，并使用自然语言情感分析技术判断文本中的情感倾向。

（3）机器学习算法

本文对比了基于情感词典的文本情感分析和基于机器学习的文本分类算法，后者的准确率明显高于前者，因此最终选择朴素贝叶斯算法（Naïve Bayes）分析 Twitter 上文本内容的情感倾向。

（4）计量经济学方法

为了构建社交媒体情绪指数并且研究社交媒体情绪与股票市场的关系，本文使用大量计量学方法进行分析，例如相关系数分析、Ganger 因果关系检验等等。

（5）深度学习模型

传统的线性回归方法很难解决多变量或者多输入的预测问题，基于 LSTM 的循环神经网络能够很好地解决多个输入变量的问题，因此在多元时间序列预测上具有更好的表现，本文使用 LSTM 构建股票时间序列预测模型。

本文的技术路线如下图所示：

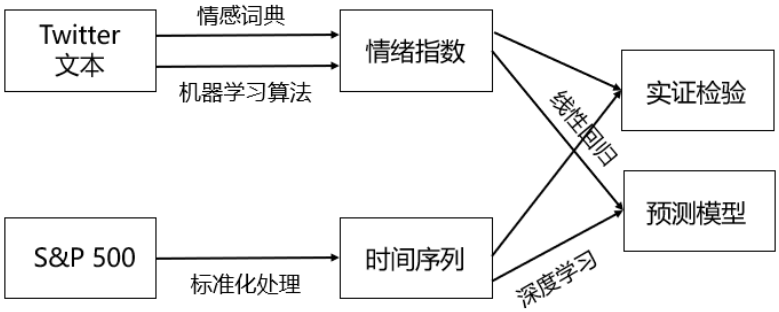

图 1 技术路线

四、情感分析

（一）数据采集与处理

1.数据来源

互联网数据研究机构 We Are Social 和 Hootsuite 共同发布的“数字 2018”互联网研究报告显示，全球超过 40 亿人使用互联网，每月超过 30 亿人使用社交媒体。社交媒体是用户发表观点和分享生活的重要媒介载体，社交媒体上的内容能够较为真实、准确地描述用户的心理状态和行为特征。全球著名的社交媒体主要包括 Facebook, Twitter，Youtube, whatsapp, LinkedIn 等，国内的社交媒体有微博，微信，QQ 等。

本文选取 Twitter 作为社交媒体数据来源。作为世界上规模最大的社交媒体，Twitter 每月的活跃用户数超过三亿，根据 Internet Usage & Social Media Statistics 的数据显示，Twitter

用户每天发布超过 5.9 亿条推文，这些推文汇聚了用户的观点、态度和情绪，因此采用 Twitter 作为情感分析的数据来源能够很好地衡量社交媒体情绪。

2020 年初社交媒体上对新冠疫情的恐慌、焦虑等情绪与世界各国股市大规模熔断之间的关联显示了社交媒体上的消极情绪对股票市场的影响。为了增加研究的拓展性，排除突发事件的影响，选取正常时期的推文更具有普遍意义。同时，由于时间过短会影响结果的准确性，因此最终本文选取 2019 年 4 月至 2019 年 9 月 6 个月的推文作为研究对象。

本文使用标普 500 指数的各项数据衡量股票市场表现，这是由于标普 500 的成分股具有涵盖范围广、代表性强、连续性好等特点，可以较为客观全面地反映股市动态。数据来源方面，本文从芝加哥期权交易所（CBOE）获取标普 500 指数的波动率（VIX），从著名财经网站 Yahoo！Finance 获得标普 500 指数的开盘价、收盘价、高低点及交易量等数据

本文主要使用编程语言 python 获取和处理数据。为了便于数据的分析和处理，本文从爬取的来自世界各地的各种语言的2,121,821,445 条推文中筛选出 1,716,602,742 条英文推文。本文使用随机函数从中挑选出 381,467,276 条数据进行后续的分析和处理，在整体的基础上压缩了 77%的工作量。

2.文本清洗与处理

文本清洗与处理是情感分析重要的前置流程，原始数据中的噪声会对分析过程造成干扰并影响结果的准确性。由于通过网络爬虫获取的 twitter 日志格式复杂，且包含了大量噪声如标点符号、@用户名的句柄等等，这些噪声信息会影响数据分析的结果，因此有必要在情感分析之前对爬取的数据进行清洗和处理。

本文首先去除了推文中@用户名格式的句柄以及其中包含的网页链接，之后从数据集里各种语言的推文中筛选出英文推文以便于文本分析，将文本中的标点符号去除后使用自然语言工具包（nltk）将句子拆分为单词，并使用其内置的的 WordNet 词典去除停用词和无意义词语，再对清洗出来的单词进行词性标注和词形还原，最后使用词袋模型对文本进行特征提取和矢量化

自然语言工具包 nltk（Natural language toolkit）是使用 python 语言编写的开源自然语言处理（NLP）库，提供了易使用的接口和丰富的文本处理库，是 python 语言中处理自然语言的重要工具。本文主要应用 nltk 进行分词、删除停用词、词性标注、词形还原等处理文本。

词性标注（Part-of-Speech tagging 或 POS tagging)是指为分词结果中的每个单词标注正确的词性，即标注每个单词是名词、动词、形容词或其他词性的过程。词形还原指将任何形式的单词还原为单词的一般形式，如将“ate”还原为“eat”。词性标注是词形还原的基础，直接对单词进行还原准确率较低，因此本文先使用 nltk 库中的 pos_tag 方法标注单词词性，之后使用 WordNetLemmatizer 函数进行还原。

（二）基于情感词典法的情感分析

情感词典是具备情感倾向的词语的集合，基于情感词典的情感分析将文本与情感词典中被标注情感极性和强度的词语进行匹配,综合计算后得到整体文本的情感倾向。情感词典的选取至关重要,使用标注完善的情感词典是得到准确分析结果的基础。目前比较成熟和完善

的英文情感词典主要有 GI（The General Inquirer）、SentiWordNet、LIWC (Linguistic Inquiry and Word Count)等。

本文选择著名的情感分析词典 SentiWordNet 作为词典来源，SentiWordNet 对 WordNet 中的词条进行情感分类并标注出其情感倾向的权重（ESULI 和 SEBASTIANI，2005）。将情感倾向分为积极、消极和客观三大类并对其进行打分。由于一个单词可以有多种词性，每种词性的含义和用法不同，因此对单词进行词性标注可以提高 SentiWordNet 识别的准确率。

Sentiwordnet 中每个单词的数据包括词性、ID、积极得分、消极得分、同义词、语义标号及同义词含义。由于一个单词往往包含多种含义，如“good”仅仅作为名词便有 4 种含义。对单词进行词性标注后，对同一种词性下的单词得分进行加权统计，计算每个单词的情感得分公式为 $poscore = \sum_{i=1}^{n} pos_{score} / sentimentscore$，此处 n 指单词的 n 种含义。

本文首先把预处理过的推文划分为单词，将单词进行词性标注后与 SentiWordne 情感词典里的单词进行匹配。由于一个单词往往具有多种含义，SentiWordnet 中的含义按照含义的常用程度排列，为了客观地衡量单词情感，本文采用两种方式计算单词的得分，第一种是对所有含义的得分进行加权处理，第二种则是选取第一个含义的得分作为单词的得分。使用词典的内置函数计算每个单词的积极得分和消极得分，两者相减得到单词的中性得分，推文中所有单词的得分加总得到该条推文的情感分数。

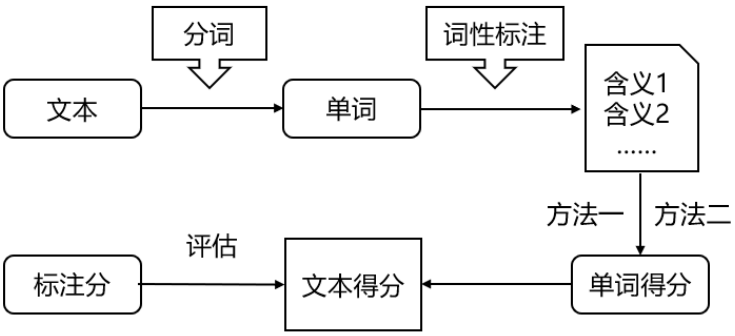

图 2 程序处理逻辑

为了评估情感词典分类的效果，本文使用 Kaggle 上已经被人工标注情感倾向的数据集 preprocessed-twitter-tweets 作为分析样本。为了方便处理，本文将该数据集中积极推文标注为 1，消极推文标注为-1，中性推文标注为 0，并将情感词典计算出的情感分数转换为与此相同的格式，从而方便对比情感词典的分类效果。

使用第一种方法，即加权一个单词所有含义的得分计算单词的得分，58%的推文情感倾向与人工标注的一致，使用第二种方法，即选取单词第一个含义的得分作为单词的得分，之后汇总得到推文的得分，67%的推文情感倾向与人工标注的一致。由此可见，使用情感词典进行情感分析准确率并不高，计算正确的推文比例甚至没有超过 70%。

由于基于情感词典的情感分析属于无监督的文本分类方法，直接将文本中的单词与情感词典中的情感词匹配从而计算情感得分，然而一个单词往往有多种含义，简单的加权计算方法容易出现较大误差。正如上文所示，无论是对含义进行加权还是直接选取靠前的含义衡量

单词，取得的分析结果都不理想，这是因为无监督的分类方法并不能准确判断单词在句子中的意思。下文将着重讨论基于机器学习的有监督的情感分类方法。

（三）基于机器学习算法的情感分析

1.文本特征表示

计算机无法理解自然语言文本，因此在使用机器学习算法处理数据前需要先将其转化为计算机能识别的数值特征。文本特征化的方式主要包括词集模型（Set of Words）和词袋模型(Bag of Words)。词集模型用 0-1 作为文本中单词数量的表示，只关注单词的有无而不关注具体数量。词袋模型在词集模型的基础上加入单词频次作为特征表示，是目前最常用的特征化文本的方式。

词袋模型(Bag of Words,简称 BoW)使用机器学习算法对文本进行特征表示和建模。特征表示指通过自然语言处理和数据挖掘技术将非结构化的文本转化为结构化的数据便于后续模型对文本的处理。词袋模型将文本看作装词的袋子,忽略文章的词序、语法和句法,不考虑上下文的关系，仅仅将文本内容看作单词或短语的组合，并按照单词出现的频次赋予其相应的权重。

词袋模型的主要流程包括分词、统计词频及向量化。分词之后统计每个词语在文本中出现的频次并作为该词的特征，之后将文本中的词语与其频次进行匹配从而实现向量化，得到单词与词频组成的特征矩阵。本文对训练集使用词袋模型对文本进行特征表示和提取，并带入算法模型进行训练，之后对模型的分类效果进行评估和比较。

2.训练算法模型

作为有监督的分类方法，机器学习对文本进行情感分析的核心步骤是使用人工标注的数据集训练算法模型。一个标注完善的数据集是模型训练准确的基础。本文采用斯坦福大学人工标注的 Sentiment140 作为训练集来源，该数据集包含从 Twitter API 采集的 1,600,000 万推文，每条推文都被标注了极性：0 为消极，2 为中性，4 为积极。

为了合理地选取训练集，本文做了以下实验探究训练集的数量对模型准确率的影响。准确率是指模型预测的结果与实际结果相同的比率。我先从 160 万条推文中随机抽选积极和消极的推文各 10,000 条，并使用逻辑回归模型训练这 20,000 条数据，之后使用训练好的模型对测试集进行预测，准确率为 73.26%。采用同样的方法，抽取积极和消极的推文各 100,000 条作为训练集预测的准确率达到 75.44%，这意味着每 10,000 条数据中后者比前者多预测对 218 条。为了提高准确率，本文将全部的 1,600,000 条推文进行了处理，并按照 8：2 的比例划分为训练集和测试集。

基于机器学习的情感分析的主要流程包括三个阶段：训练、测试和应用。在训练阶段，本文先将文本和人工标注的标签分离，并对文本进行特征表示，使其转换成为计算机能够理解的向量，之后运用算法模型对提取出的特征和标签进行拟合，从而得到基于训练集的分类器。在测试阶段，使用训练阶段得到的分类器对人工标注的测试集进行预测，将预测结果和人工标注的结果对比分析即可评估算法模型的分类效果。最后，选择测试阶段表现最好的算法模型应用到实际数据集中，对数据集的情感倾向做出预测。

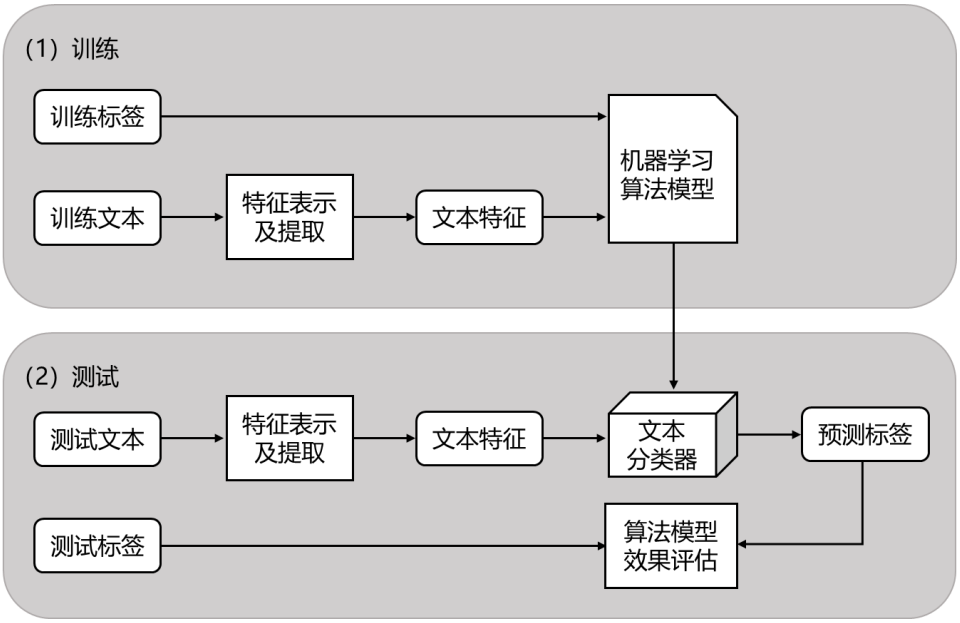

图 3 机器学习流程

由于分类算法的理论基础和实现原理不同，不同的分类算法适用于不同的领域和场景，不同的数据集上分类算法的表现也会有差别。因此为了选出适合对 Twitter 数据集进行情感分析的算法，本文使用支持向量机（SVM）、朴素贝叶斯模型（NB）、K 近邻算法（KNN）、Logistic 回归模型（LR）和随机森林（RF）五种机器学习算法对训练集进行拟合并评估每个算法模型的分类效果，从而选择最适合本文的算法模型。

本文使用以下四个机器学习领域常用的衡量指标对算法的文本分类效果进行评估：

准确率 Accuracy=预测正确的样本数除以总样本数；

精准度 Precision=预测为正例的样本数除以实际为正例的样本数；

召回率 Recall=总样本中的正例数除以预测正确的样本数；

F1 值 F1-score = 2 ×Precision×Recall/（Precision + Recall）

为了测试各个分类器的训练效果，本文先使用词袋模型对文本进行特征表示，之后采用相同的 20000 条训练集和 500 条测试集分别对算法模型进行训练和测试，评估结果如下表所示：

表 1 基于词袋模型的分类算法评估结果

	Accuracy	Precision		Recall		F1	
		neg	pos	neg	pos	neg	pos
SVM	0.76	0.80	0.74	0.70	0.83	0.75	0.78
NBM	0.79	0.79	0.78	0.77	0.80	0.78	0.79
KNN	0.65	0.72	0.62	0.49	0.81	0.58	0.70
RF	0.75	0.78	0.73	0.69	0.81	0.73	0.77
LR	0.78	0.80	0.75	0.72	0.83	0.76	0.79

从上表可以看出，K 近邻算法的表现最差，朴素贝叶斯和 Logistic 回归模型的表现最好。相较于 Logistic 回归模型，朴素贝叶斯对积极和消极的分类效果更均衡，因此本文最终采用

朴素贝叶斯模型对文本进行情感分类。

表 2 训练后的朴素贝叶斯模型效果

	Accuracy	Precision	Recall	F1-score
neg	0.88	0.88	0.88	0.88
pos	0.88	0.89	0.88	0.89

选取算法模型后，本文使用 Sentiment140 中的 160 万条标注推文对算法模型进行训练，最终的情感分类器准确率高达 88%，远远超过情感词典的 67%。情感分析的效果直接影响到实证研究的准确性，因此本文最终采用机器学习方法对进行情感分析。

（四）构建社交媒体情绪指数

在对获取到的文本进行特征表示后，本文使用贝叶斯模型对获取到的推文进行分类，从而判断推文的情感倾向。使用机器机器学习算法实现情感分类后，分别统计积极推文数量和消极推文数量，并以此为基础构建社交媒体情绪指数。

由于本文的研究目的为探究社交媒体和股票市场表现之间的关系，因此在构建指数时本文借鉴了行为金融学领域的相关研究，如使用 Werner Antweiler 和 Murray Z. Frank 构建投资者情绪指标[39]的方法构建看涨指数，借鉴 Sanjiv R. Das 和 Mike Y. Chen 的方法构建情感分歧指数[40]等等。为了从多个角度描述社交媒体情绪，本文构建以下社交媒体情绪指标。

简单情绪指数（Simple Sentiment index, SSI）：

$$SSI_t = N_t^{pos} - N_t^{neg}$$

积极情绪指数（Positive Sentiment index, PSI）

$$PSI_t = \frac{N_t^{pos}}{N_t^{pos} + N_t^{neg}}$$

消极情绪指数（Negative Sentiment index, NSI）

$$NSI_t = \frac{N_t^{neg}}{N_t^{pos} + N_t^{neg}}$$

看涨指数（Bullishness Sentiment index，BSI）

$$BSI_t = \ln\left[\frac{1 + N_t^{pos}}{1 + N_t^{neg}}\right]$$

情感分歧指数(Distinguish index sentiment, DIS)

$$DIS_t = 1 - \left|\frac{N_t^{pos} - N_t^{neg}}{N_t^{pos} + N_t^{neg}}\right|$$

其中，N_t^{pos}代表 t 时间段内积极情绪推文的数量，N_t^{neg}代表 t 时间段内消极情绪推文的数量。本文通过机器学习方法标注情感后，利用上述公式计算相关指数，并构建社交媒体情绪指数时间序列。

五、实证研究

（一）相关性分析

相关性分析是一种研究变量之间相关关系的统计分析方法，常用的相关分析方法包括图表分析法、协方差矩阵法、相关系数法、回归分析法等等。由于协方差适用于定性分析，即相关性的正负和有无，无法衡量相关性的大小，因此本文采用相关系数法探究各项指标之间的相关性。

统计学中常用的相关系数包括皮尔森（Pearson）相关系数、斯皮尔曼（Spearman）相关系数和肯德尔（Kendall）相关系数。与 Pearson 相关系数不同，Spearman 相关系数和 Kendall 相关系数使用等级而非具体值反应变量间的相关程度。为了更细致地衡量相关性，本文使用 Pearson 相关系数对变量进行两两之间的相关性检验。

皮尔森相关性系数是用来衡量变量之间线性关联度的常用方法，计算公式为变量 x,y 之间的协方差除以各自标准差的乘积，即：

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

皮尔森相关系数的取值在-1 和 1 之间，相关系数的绝对值越大，相关性越强，相关系数为 0 时代表变量之间没有关联。由于消极情感指数、情感分歧指数和积极情感指数为负线性相关，看涨指数和积极情感指数为正线性相关，因此选取积极情感指数即可代表消极情感指数、情感分歧指数和看涨指数与股票市场指标进行相关性分析。SPSS 计算各指标之间的皮尔森相关系数结果如下：

表 3 相关性分析

	Volume	收益率	VIX	PSI	NSI	SSI	BSI
Volume	1						
收益率	-0.186*	1					
VIX	0.295**	-0.299**	1				
PSI	0.034	-0.001	-0.302**	1			
NSI	-0.034	0.001	0.302**	-1.000**	1		
SSI	0.037	-0.015	-0.283**	0.867**	-0.867**	1	
BSI	0.034	-0.001	-0.302**	1.000**	-1.000**	0.868**	1
DIS	-0.034	0.001	0.302**	-1.000**	1.000**	-0.867**	-1.000**

**. 在 0.01 级别（双尾），相关性显著。

*. 在 0.05 级别（双尾），相关性显著。

由相关性分析结果可以看出，积极情绪指数（PSI）、简单情绪指数（SSI）和情感分歧指数（DIS）和标普 500 指数波动率（VIX）呈显著负相关，积极情绪指数上升时，标普 500 指数波动率下降；情感分歧指数上升时，即社交媒体中消极或积极的情绪较为分散时，标普 500 指数波动率上升，这也与事实相符。

（二）VAR 模型和 Granger 因果关系检验

VAR 模型（向量自回归模型）是一种用来估计联合内生变量之间动态关系的计量经济

模型，常用于处理多个相关经济指标的分析和预测。VAR 模型是对标准的自回归模型（AR 模型）的拓展，从单一时间序列拓展到多元时间序列。在 VAR 模型中，同一样本期的变量对所有变量的滞后项进行回归从而构造模型，一个 p 阶的 VAR 模型可以表示为：

$$y_t = \sum_{k=1}^{d} A_k y_{t-k} + \epsilon_t, t = d+1, \ldots, T$$

其中 k=1,2,…,d 表示向量自回归的系数矩阵，ϵ_t可视为高斯噪声。

构建 VAR 模型的前提是平稳时间序列，平稳性检验方法的常用方法包括 DF 检验、ADF 检验和 PP 检验等，本文使用 ADF 检验对相关指标进行平稳性检验，检验结果及标准如下：

表 4 ADF 检验标准值

置信水平	临界值
1%	-3.513
5%	-2.892
10%	-2.581

表 5 标普 500 指数 ADF 检验

指标	t	p-value	是否平稳
开盘价	-1.775	0.3928	否
收盘价	-2.524	0.1096	否
交易量	-4.789	0.0001	是
波动率	-3.375	0.0106	是
收益率	-11.722	0.0000	是

表 6 社交媒体情绪指数 ADF 检验

指标	t	p-value	是否平稳
积极情绪指数（PSI）	-6.815	0.0000	是
消极情绪指数（NSI）	-6.815	0.0000	是
简单情绪指数（SSI）	-7.414	0.0000	是
看涨指数（BSI）	-6.838	0.0000	是
情感分歧指数（DIS）	-6.815	0.0000	是

由以上表格可以看出，对标普 500 指数进行 ADF 检验，开盘价和收盘价时间序列不是平稳序列，交易量和收益率时间序列在 1%的置信水平上平稳，波动率时间序列在 5%的置信水平上平稳。对社交媒体情绪指数进行 ADF 检验，PSI、NSI、SSI、BSI、DIS 时间序列均在 1%的置信水平上平稳。

本文对检验平稳的时间序列 Volume、收益率、VIX、PSI、BSI、SSI 通过 z-score 标准化处理后构建 VAR 模型，并使用 Granger 因果关系检验和脉冲响应函数对 VAR 结果进行评价和分析。Granger 因果关系检验（Granger Causal Relation Test）是用于分析经济学变量之间

Granger 因果关系的常用方法，其统计学本质是对平稳时间序列的预测。

本文先根据 AIC 和 SC 取值最小准则确定 Granger 因果关系检验的最佳滞后阶数，如果 AIC 和 SC 并不是同时取值最小，则根据 LR 检验结果的最小值选择阶数。一般采用 4 阶作为最大滞后阶数，滞后阶数超过 4 阶检验出的因果关系往往没有意义。

零假设为H_0：X dose not Granger-cause Y，在 95%的置信水平上对变量进行 Granger 因果关系检验，结果如下：

表 7 Granger 因果关系检验结果

Y	X	Chi2	df	Prob>chi2
VIX	PSI	16.888	4	0.002
	BSI	16.469	4	0.000
	SSI	63.312	4	0.000
PSI	VIX	90.092	4	0.000
	Volume	53.619	4	0.000
	收益率	72.923	4	0.000
Volume	PSI	103.930	4	0.000
	BSI	103.690	4	0.000
	SSI	66.770	4	0.000
BSI	VIX	88.461	4	0.000
	Volume	52.760	4	0.000
	收益率	480.090	4	0.000
收益率	PSI	47.718	4	0.000
	BSI	47.445	4	0.000
	SSI	31.649	4	0.000
SSI	VIX	50.604	4	0.000
	Volume	42.235	4	0.000
	收益率	41.999	4	0.000

由于 NSI、DIS 与 PSI 线性相关所以 Granger 因果检验的结果与 PSI 相同，p 值为 0.000。结合表中数据可得，各变量 Granger 因果检验的 p 值均小于 0.05，因此社交媒体情绪指数在 95%的置信水平上与标普 500 指数波动率、交易量和收益率互为 Granger 因果关系。这表明，加入社交媒体情绪指数有助于预测标普 500 指数，标普 500 指数的变动也有助于解释社交媒体情绪指数的变化。

（三）脉冲响应分析

Granger 因果关系检验描述了变量之间的因果关系，但仅能说明一个变量是否有助于解释和预测其他变量，并不能判定变量之间的作用方向及影响时间。脉冲响应分析通过对 VAR 模型中的变量施加“外生冲击”，观察模型中其他变量受到的动态影响，从而判断变量之间的变化趋势。

本文先对变量使用 VAR 模型拟合，之后使用脉冲响应函数预测 8 期的变动，脉冲响应函数用于衡量随机扰动项的一个标准差冲击对内生变量的影响。由于脉冲响应函数中因变量都会被顺序在其之前变量的冲击影响，而不会被顺序在其之后的变量影响，为了避免这种干扰，本文尽量减少函数中的变量，根据变量受到冲击后一段时间内的动态变化绘制出脉冲响应图形，以下为两幅示意图：

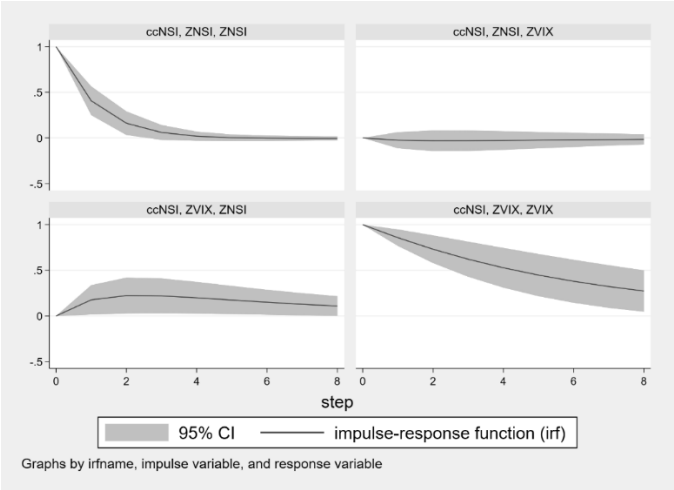

图 4 NSI-VIX 脉冲响应图

脉冲响应图中每行放置不同变量受同一变量冲击的影响，每列放置同一变量受不同变量冲击的影响。横轴为预测期数，上图中为 8 期，即冲击在未来 8 期内带来的影响。

由 NSI-VIX 脉冲响应图可以看出，VIX 和 NSI 之间的影响集中在前两期，随后逐渐收敛，波动率的冲击会引起消极情绪指数的显著上涨，这也与实际情况相符。

本文对三个标普 500 指数变量和五个社交媒体情绪变量进行两两脉冲响应图分析，最终得出变量之间的相互影响关系。如上两张示意图所示，变量之间的影响关系大多集中在前两期，即冲击带来的影响往往是短暂的，并不会持续很久。这也与社交媒体和标普 500 指数时效性强的特性相符。

（四）多元线性回归模型预测

为了研究社交媒体情绪指数对于标普 500 指数的解释和预测能力，本文使用 python 对社交媒体情绪指数和标普 500 指数进行回归分析。由上文可知标普 500 指数的指标中收益率、交易量和波动率是平稳序列，社交媒体情绪指数中 PSI、BSI 和 SSI 是不具备线性关系的平稳序列，因此分别将收益率、交易量和波动率作为自变量，其他指标作为因变量进行回归。

本文首先将所有数据划分为训练集和测试集，并使用线性回归模型拟合训练集，之后用训练好的线性回归模型对测试集进行预测。线性回归模型假设变量间存在线性相关关系，并根据各变量权重赋予回归系数，之后结合截距和回归系数计算出线性回归方程。

以收益率预测为例，先不加入社交媒体情绪指数进行回归分析。通过线性回归模型计算出截距为 0.027，各变量的系数为 open：-2.503，high：-0.060，low：0.327，close：2.264,则其线性回归方程为:

$$收益率 = 0.027 - 2.503\text{open} - 0.060\text{high} + 0.327\text{low} + 2.64\text{close}$$

预测值和实际值对比如下：

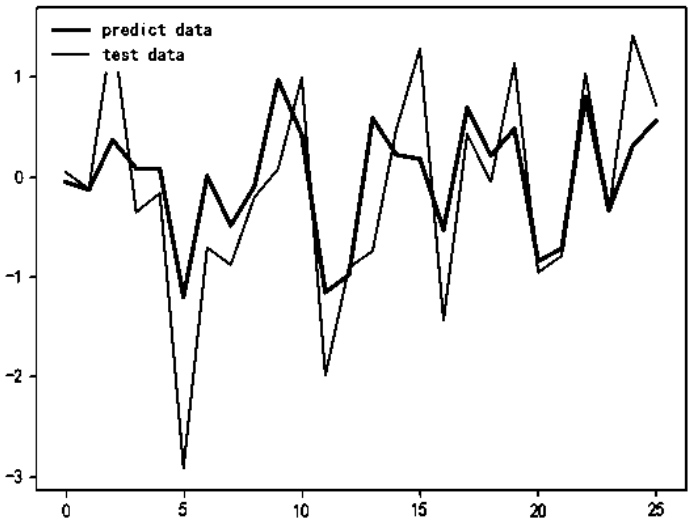

图 5 多元线性回归预测收益率与实际收益率对比

本文使用均方根误差（Root Mean Squared Error）对回归结果进行评估，该回归方程均方根误差 0.4435，加入 PSI 和 NSI 后回归方程均方根误差为 0.4302，均方根误差越小代表拟合度越高，因此社交媒体情绪指数有助于对标普 500 指数进行解释和预测。

（五）基于 LSTM 神经网络的时间序列预测

LSTM 长短期记忆网络（ Long-Short Term Memory networks）是一种时间递归神经网络，也是 RNN 循环神经网络（Recurrent Neural Network）的一个变形。LSTM 的设计结构是为了解决一般的 RNN 存在的长期依赖问题，因此相较于 RNN，LSTM 更适合处理和预测间隔和延迟长的时间序列。

传统的线性回归方法很难解决多变量或者多输入的预测问题，基于 LSTM 的循环神经网络能够很好地解决多个输入变量的问题，因此在多元时间序列预测上具有更好的表现。作为典型的非线性模型，LSTM 也可以作为复杂的非线性单元构造更大型的深度神经网络。

时间序列预测分析是指利用过去一段时间变量的取值预测其未来取值，是一类比较复杂的预测建模问题。不同于简单的回归分析，时间序列预测依赖于变量值的先后顺序，调换变量值的顺序则预测结果也会完全改变。

由于多元线性回归模型的预测效果并不理想，因此本文采用非线性的 LSTM 模型对股票数据进行预测。为了便于与多元线性回归模型的预测结果对比，本文使用相同的数据集，并采用相同的标准化方法对数据进行处理。

本文使用 LSTM 模型隐藏层的 50 个神经元对数据进行深度学习，并在训练和测试完成后使用 MAE（平均绝对误差）作为损失函数绘制损失曲线。损失函数用于描述预测值和真实值之间的差距，损失值越小则预测值越接近真实值。

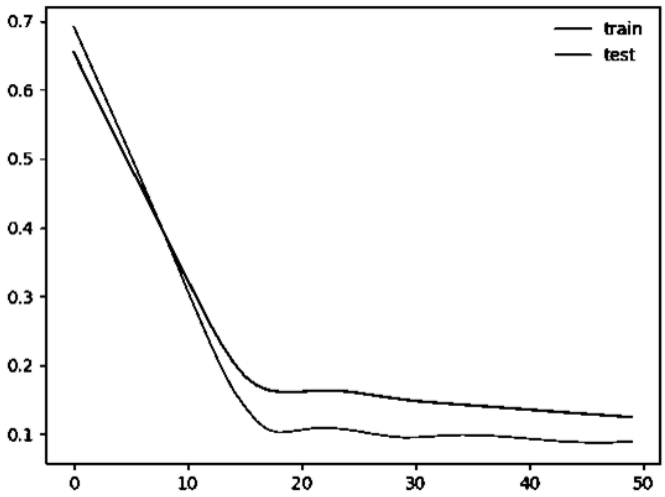

图 4-8 训练集和测试集的损失曲线

如上图所示，随着神经元数目增加，训练集和测试集的损失先下降后收敛。这是由于前期模型仍在学习，因此损失逐渐下跌。后期训练损失和测试损失逐渐收敛并趋于 0，表明模型已经停止训练。本文的测试损失小于训练损失，说明 LSTM 模型可能过度拟合训练集。

与多元线性回归模型衡量方式相同，采用均方根误差（RMSE）对 LSTM 模型预测结果进行评估。当自变量为开盘价、收盘价、高点、低点，因变量是收益率时，均方根误差为 0.154，远远低于多元线性回归模型的均方根误差。将 PSI、NSI 和 BSI 加入自变量后均方根误差下降至 0.120，这表明社交媒体情绪有助于预测收益率时间序列。

六、研究结论与启示

资产定价一直是金融学的核心问题，经典金融学在均衡定价和无套利定价的理论框架下构建出一系列资产定价模型，行为金融学在经典金融学的基础上融入行为人的心理因素，对原有的模型加以修正和完善。社交媒体作为人们表达观点和分享生活的载体，包含了巨量的用户心理和行为特征。结合社交媒体上的大数据和行为金融学理论分析股票市场走势不仅可以充分发挥社交媒体信息丰富的优势，还能够有效地探究公众情绪对股票市场的影响。

本文选取 2019 年 4 月 1 日至 9 月 30 日的 Twitter 文本和标普 500 指数作为数据来源，以行为金融学理论为基础，应用机器学习方法和计量经济学模型对数据进行分析和处理，得出以下结论：

第一，本文分别使用情感词典方法和多种机器学习算法对获取到的 Twitter 文本进行文本分类和情感标注，使用相同的测试集评估情感分析结果，基于机器学习的情感分析平均准确率为 75%，远远高于基于情感词典的情感分析准确率 67%。测试的五种机器学习算法中，朴素贝叶斯模型的表现最好，训练后的朴素贝叶斯模型准确率高达 88%，因此本文采用朴素贝叶斯模型进行文本分类。

第二，本文借鉴行为金融学领域的相关研究构建社交媒体情绪指数，采用多种实证研究方法探究社交媒体情绪指数和股票市场的关系。关联性分析中，本文使用皮尔森相关系数

衡量变量之间的线性关系，发现社交媒体情绪指数和标普 500 指数的波动率显著相关； 动态关系分析中，本文采用 VAR 模型、Granger 因果关系检验、脉冲响应分析等多种计量经济学方法，结果表明社交媒体情绪指数有助于解释和预测标普 500 指数的变动，但是两者只存在短期的影响关系。

第三，本文分别使用多元线性回归模型和 LSTM 神经网络对时间序列进行预测，两者的预测结果均表明加入社交媒体情绪指数的模型误差更小、拟合度更高，因此社交媒体情绪指数有助于解释和预测股票市场的表现。相较于多元线性回归模型，基于 LSTM 神经网络的预测效果更好，因此各变量间可能存在着较强的非线性关系。

参考文献

[1] Froot K A, Scharfstein D S, Stein J C. Herd on the Street: Informational Inefficiencies in A Market with Short-Term Speculation[J]. National Bureau of Economic Research Cambridge, Mass. USA. 1990.

[2] Data source: Internet Live Stats - Internet Usage & Social Media Statistics

[3] Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk[J]. Econometrica: Journal of the Econometric Society, 1997: 263-291.

[4] Osborn H, Engineers S.o.A.Latest Developments in High Frequency Welding. Society of Automotive Engineers,1964

[5] Markowitz Harry M. Portfolio selection. Journal of Finance,1952,7(1),77-91

[6] Ross J, Eady E, Cove J,et al. Inducible erythromycin resistance in staphylococci is encoded by a member of the ATP-binding transport super-gene family. Molecular microbiology,1990,4(7),1207-1214

[7] De Bondt W F M. Does the stock market overreact to new information? /[J]. Journal of Finance, 1985, 40(3):793-805.

[8] Mehra R. and E.C. Prescott. The Equity Premium: A Puzzle[J].Journal of Monetary Economics, 1985,15(2):145-161.

[9] Baker, Malcolm, Jeffrey Wurgler. Journal of Economic Perspectives 21, no. 2 (Spring 2007): 129–151.

[10] DE L J B，SHLEIFE R A，SUMME R S L H，et al． Noise trader risk in financial markets［J］． J Polit Econ，1990，98: 703- 738．

[11] BLACK F． Noise［J］． J Fin，1986，41(3) : 528-543．

[12] D. Gruhl, R. Guha, R. Kumar, J. Novak, A. Tomkins, The predictive power of online chatter, in: KDD ‘05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM Press, New York, NY, 2005, pp. 78–87.

[13] G. Mishne, M.D. Rijke, Capturing global mood levels using blog posts, in: N.Nicolov, F. Salvetti, M. Liberman, J.H. Martin (Eds.), AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, The AAAI Press, Menlo Park, CA/Stanford University, CA, August, 2006, pp. 145–152.

[14] H. Choi, H. Varian, Predicting the Present with Google Trends, Tech. rep., Google,2009.

[15] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news, ACM Transactions on Information Systems 27 (February (2)) (2009) 1–19

[16] Karabulut Yigitcan, Can Facebook Predict Stock Market Activity?, SSRN Electronic Journal,

DOI: 10.2139/ssrn.2017099
[17] Johan Bollena,Huina Maoa,Xiaojun Zeng,Twitter mood predicts the stock market, Journal of computational science,2011,2(1):1-
[18] Kahneman, D. and A. Tversky, 1979,"Prospect Theory: An Analysis of Decision under Risk", Econometrica, 47, pp.263~291.
[19] Banerjee， A. V. , A Simple Model of Herd Behavior, Quarterly Journal of Economics 107(3):797-817
[20] Sullivan D. The Need for Text Mining in Business Intelligence. Published in DM Review in Dec.2000
[21]MACDONALD C，OUNIS I． The TREC Blogs 06 collection: Creating and analysing a blog test collection［R］. Glasgow，Scotland: University of Glasgow， Department of Computer Science，2006．
[22]SEKI Y,EVANS D K,KU L W，et .al. Overview of opinion analysis pilot task at NTCIR-6［C］/ / Proceedings of the Workshop Meeting of the National Institute of Informatics Test Collection for Information Retrieval Systems． Tokyo: National Center of Science，2007: 265 － 278．
[23] Salton G, Wong A, Yang C S.A Vector Space Model for Automatic Indexing. Communication of the ACM 1995,18:613-620
[24] Quinlan J R. Constructing Decision Tree in C4.5. Morgan Kaufman Publishers, Programs for Machine Learing,1993.17-26
[25] Yang Y, Pedersen J O.A Comparative Study on Feature Selection in Text Categorization .In : Proc. of the 14th Intl. Conf. on Machine Learning ICML 97,1997.412-420
[26] Jolliffe I T . Principal component analysis[M] .2nd ed. New York: Springer , 2002.
[27] McCallum A, Nigam K.A Comparison of Event Models For Naïve Bayes Text Classification. Just Research 4616 Henry Street Pittsburgh, PA 15213
[28] Mladenic D. Machine Learning On Non-Homogeneous, Distributed Text Data. Doctoral Dissertation, University of Ljubljana,1998
[29]NASUKAWAT,YIJ. Sentiment analysis : Capturing favorability using natural language processing [C]//Proceedings of the 2nd International Conference on Knowledge Capture. New York: ACM Press 2003:70-77.
[30]HU M， LIU B． Mining opinion features in customer reviews［C］ / /Proceedings of the National Conference on Artificial Intelligence．San Jose， California: AAAI Press，2004: 755 － 760．
[31] STONE P． The general inquirer: A computer approach to content analysis［J］. Journal of Regional Science，1968，8(1) : 113 － 116．
[32] ESULI A， SEBASTIANI F． Determining the semantic orientation of terms through gloss analysis［C］ / / Proceedings of the ACM SIGIR Conference on Information and Knowledge Management． New York: ACM，2005: 617 － 624．
[33] PANG B， LEE L， VAITHYANATHAN S. Thumbs up： sentiment classification using machine learning techniques[C]. Acl-02 Conference on Empirical Methods in Natural Language Processing，2002：79-86.
[34] WILSON T， HOFFMANN P， SOMASUNDARAN S， et al．OpinionFinder: A system for subjectivity analysis ［ C ］/ /Proceedings of HLT / EMNLP on Interactive Demonstrations．Stroudsburg， PA: Association for Computational Linguistics，2005: 34 － 35．
[35] GAMON M， AUE A， CORSTON-OLIVER S， et al. Pulse: Mining customer opinions

from free text［C］/ / Proceedings of the6th International Symposium on Intelligent Data Analysis． Berlin: Springer-Verlag，2005: 121 － 132．
[36] Hart，R． P． 2000． DICTION 5． 0: The Text-analysis Program［M］． Thousand Oaks，CA: Sage．
[37] Vapnik, V.. Statistical learning theory. 1998 (Vol. 3). . New York, NY：Wiley，1998：Chapter 10-11, pp.401-492
[38]Hsieh, W.W.. Machine learning methods in the environmental sciences: Neural networks and kernels：Cambridge university press，2009：Chapter 7, pp.157-169
[39] Antweiler W, Frank M Z. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards[J]. The Journal of Finance, 2004, 59(3):1259–1294.
[40] Das S R, Chen M Y. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web[J]. Management Science, 2007, 53(9):1375-1388.

The impact of social media on the stock market

——Empirical analysis based on Text Mining

Abstract: The performance of stock market not only reflects a country's economic development and capital market trend, but also affects individual investment decision, company's business strategy, and even the national economic policy. This paper uses big data on social media to analyze and predict the stock market, and explores the impact of social media sentiment on the stock market with the help of text mining techniques, machine learning algorithms, and econometric methods. This article constructs social media sentiment index from multiple dimensions, and conduct empirical research on the performance of the stock market. The results show that there is a strong non-linear correlation between social media sentiment and stock market. Joining social media sentiment index helps to explain and predict the stock market.

Keywords: text mining; sentiment analysis; machine learning; social media; stock market