基于 Text-CNN 模型的散户投资者情绪指数与股指关系研究

李珊[1]　丁宇[1]　何呈琳[1]　刘晓星[2*]

（1.南京航空航天大学经济管理学院，南京，211000）

（2.东南大学经济管理学院，南京，211189）

摘要：我国的股票市场以散户投资者为主，散户投资者的情绪对股指有着重要影响。为了研究我国散户投资者情绪与股指之间的相互关系，本文以网股吧论坛散户发帖内容为样本数据，利用 Text-CNN 模型对发帖内容进行情感分类，并在此基础上构建投资者情绪指数，然后对投资者情绪指数和上证股指进行格兰杰因果关系分析、脉冲响应分析和方差分解分析。研究结果显示：散户投资者的积极情绪与股指之间存在着长期且正向的相互作用，并且这种作用在不受其他因素干扰时呈现出逐渐增强的趋势。在没有利空消息披露和重大风险形成时，散户投资者的积极情绪和股指对彼此会施加一个长期的不断增强的促进作用，导致散户的投资热情不断升温和股票指数的不断攀高。

关键词：散户投资者；Text-CNN 模型；文本处理；投资者情绪指数；上证股指

Research on Relationship between Stock Index and Retail Investor Sentiment Index Based on Text-CNN Model

Li Shan[1]　Ding Yu[1]　He Chenlin[1]　Liu Xiaoxing[2*]

(1.Nanjing University of Aeronautics and Astronautics,School of Economics and Management,NanJing,211000)

(2.Southeast University,School of Economics and Management,211189)

Abstract: China's stock market is dominated by retail investors, and the sentiment of retail investors has an important impact on the stock index.In this paper, the content of retail posts of Dongfang fortune net is taken as the sample data, and the Text-CNN model is used to classify the emotion of the posts. On this basis, we constructed the investor sentiment index, and then granger causality analysis, impulse response analysis and variance decomposition analysis are conducted to find the relationship betweeninvestor sentiment index and Shanghai stock index.The results show that there is a long-term and positive interaction between the positive sentiment of retail investors and the stock index, and this kind of interaction shows a trend of gradually strengthening as it is not disturbed by other factors.This shows that in the absence of bad news disclosure and significant risk formation, the positive sentiment and stock index of retail investors will exert a long-term and increasing promotion effect on each other, leading to the rising investment enthusiasm of retail investors and the rising stock index.

Keywords Retail investors; Text – CNN model; Text processing; Investor sentiment index; The Shanghai stock index

李珊(1977-),女，汉族,副教授，研究方向：数据处理、金融智能，lishan@nuaa.edu.cn

丁宇（1996-）男，汉族，硕士生，研究方向：数据分析，金融智能，931415288@qq.com

何呈琳（1995-）,女，汉族，硕士生，研究方向：金融风险，互联网金融，hechenglin@nuaa.edu.cn

刘晓星(1970-),男，汉族,教授，通讯作者，研究方向：金融工程、金融智能，starsunmoon198@163.com

国家社科基金重大专项课题(No.18VSJ035),国家自然科学基金项目(No.71673043),中央高效基本科研业务费专项资金（No. NJ2019023），国家社会科学基金项目(17BGL055) ，南京航空航天大学研究生创新基地（实验室）开放基金项目(kfjj20180910)资助。

一、引言

投资者情绪是行为金融学研究的一个热点。行为金融学理论认为证券的价格不只是由其内在的价值所决定，投资者的投资情绪和行为也会对其产生很大的影响。De Long 等(1990)最先研究投资者情绪与股票价格的关系，并构建了 DSSW 模型，指出投资者情绪是影响金融资产均衡价格的系统性风险[1]。通过构建投资者情绪指数来对投资者的情绪进行量化分析是目前主要的研究方式。随着互联网的快速发展，海量的互联网数据和日趋成熟的数据处理技术为研究投资者情绪提供了新的思路和方法。Antweiler(2004)利用贝叶斯和支持向量机的方法对雅虎金融板块上近 150 万条数据进行分析并构建投资者情绪指数，研究发现股票收益率与当日论坛中的信息指标呈显著的相关关系[2]。Tetlock(2007)通过对媒体新闻内容和股票收益率进行研究，发现悲观的新闻报道会对股票收益率产生影响[3]。孟雪井等(2016) 利用文本挖掘技术建立了三种词库，并在综合三种词库基础上构建了投资者情绪指数，证明了我国股票市场的投资者情绪与市场指数之间存在联动机制[4]。石善冲等(2018)基于微信文本构建了投资者情绪指数，并与收盘价、成交量时间序列之间的关系进行研究，结果表明于微信文本挖掘的投资者情绪对于预测股票市场表现有重要作用[5]。

基于互联网数据构建情绪指数的基本思路是：通过爬虫技术获取大量互联网文本数据，通过本文分析方法进行情感分类，在此基础上构建投资者情绪指数，再对其进行相关分析。目前用于构建投资者情绪指数的文本分析方法主要是：机器学习法和情感词典法[6-9]。机器学习法主要利用传统的机器学习算法包括决策树、朴素贝叶斯和支持向量机等，需要选取或者建立文本特征，然后进行情感分类，此方法的关键在于选择合理的特征。情感词典法需要先建立相关领域文本的情感词典，在此基础上对各文本数据进行情感打分，最后根据得分进行情感分类。机器学习算法其效果很大程度上取决于文本特征的选取[10]；情感词典法目前在金融领域没有一个能完整包含所有词语语义倾向的词典，而且词语在不同语境中所表现出得不同情感倾向问题造成情感分类的不准确[11]。随着对深度学习领域的不断探索和研究，越来越多的学者开始利用深度学习算法来进行文本分析处理[12-15]。梁军等(2014)提出了一种基于深度学习的递归神经网络模型对微博文本情感进行分析，在不增加神经网络复杂度的前提下，提高了模型分析的准确性[15]。李杰等(2018) 采用卷积神经网络模型对电商评论进行短文本评论情感分类和特征词聚类，结果表明模型的情感分类达到高准确率，并且能有效提取产品特征[16]。基于深度学习的文本处理算法正不断受到各领域学者的青睐，然而在对投资者情绪指数研究上，还没有引起广泛关注。

本文将深度学习算法引入投资者情绪的研究中。首先选取我国散户投资者为研究对象，

获取股吧论坛评论文本，以基于卷积神经网络的文本分类算法(Text-CNN)来对投资者情感进行分类；然后在此基础上构建投资者情绪指数，最后进行实证，揭示投资者情绪指数与上证股指的相互关系。

二、模型和研究设计

(一)Text-CNN 模型

自然语言处理一直是学术界研究的热点，随着对深度学习研究的不断深入和创新，Kim（2014）首次将自然语言处理和深度学习中的卷积神经网络(CNN)相结合，提出了基于卷积神经网络的文本分类模型(Text-CNN)[17], Zhang Y等（2015）对Text-CNN模型具体的调参细节进行了详尽的阐述[18]。该模型的基本结构是在传统卷积神经网络模型的卷积层之前添加一个嵌入层，从而实现自然语言到为计算机数字语言的转换，同时设置窗口大小不同的三种卷积核对句子进行一维卷积，随后在池化层中使用一维1-MaxPooling对卷积的结果进行池化，并将最终的池化结果进行拼接，将池化拼接后的结果连接到全连接层，进行分类。在基本的Text-CNN模型中，嵌入层和其它层一样，所有的参数都是随机初始化，然后通过BP算法不断进行调整，最后经过样本训练得到模型内部的所有参数。图1展示了一个典型的双卷积核Text-CNN模型结构图。

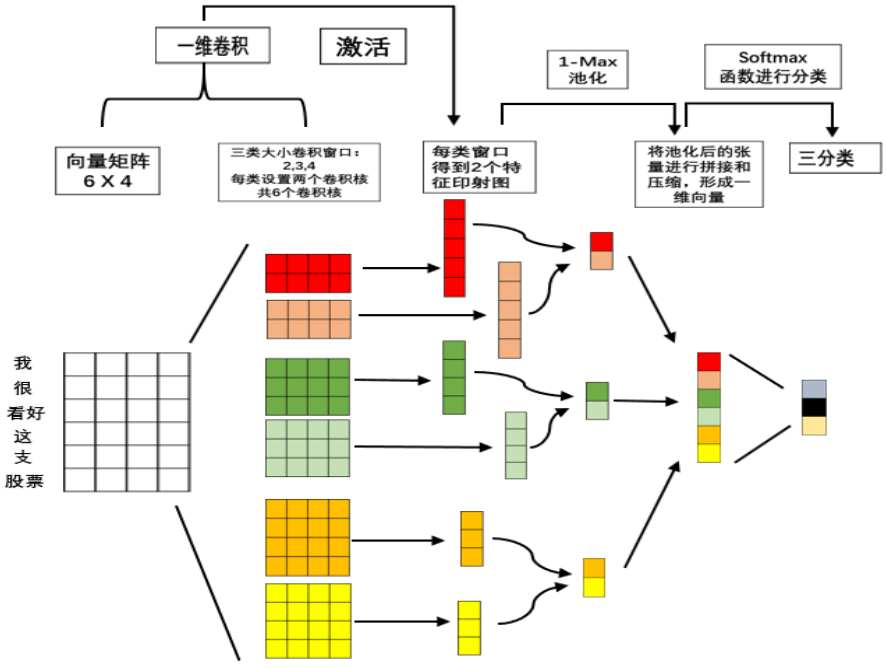


图1 双卷积核Text-CNN模型结构示意图

(二)投资者情绪指数设计

选取某一股市论坛帖子作为数据源，利用构建的Text-CNN模型对收集的散户投资者发

帖进行情感分类并获得每条样本分类的概率值(p)，同时计算每条帖子的关注度指数(fi, focus index)。关注度指数是用于衡量某条帖子所受关注的程度，关注度指数与该帖受到的关注度成正比，指数越高说明该帖的内容和情感在散户投资者中传播的范围越广，影响越大。选取帖子的“阅读数(qr, quantity of read)”和“评论数(qc, quantity of comment)”作为指标，正常情况下，用户会对感兴趣帖子的内容点击阅读，同时只会在对帖子产生强烈兴趣时评论帖子，说明“评论数”所代表的关注度要高于“阅读数”，因此在构建关注度指数时，需要提高“评论数”的权重，本文将qc权重设置为5，qr权重为1。每条帖子关注度指数的公式是：

$$fi = qr + 5qc \tag{1}$$

股市成交量为股票买卖双方达成交易的数量，成交量对股票价格有着重要影响，研究表明：不管是个股还是证券组合，价格波动与成交量之间呈正相关关系[19]。由于投资者的投资心理和实际投资行为可能会产生偏差，例如：表达积极情绪，但是可能因为各种原因采取继续观望的策略，同时存在部分投资者只浏览他人帖子但自己不发表意见，而成交量指标能在一定程度上反应所有投资者投资行为的积极与否。考虑上述原因，本文将股市每日成交量指标添加到投资者关注度指数中，对投资者情绪指标进行修正。

投资者情绪指数由每条帖子关注度指数(fi)、分类概率值(p)以及每日成交量（smt, stock market turnover）组成，为了对积极情绪和消极情绪分别展开研究，构建两类投资者情绪指数：积极投资者情绪指数(PSI)和消极投资者情绪指数(NSI)，指数计算公式如下:

$$PSI = norm(\sum(fi_1 \times p_1)) + norm(smt) \tag{2}$$

$$NSI = -norm(\sum(fi_2 \times p_2)) + norm(smt) \tag{3}$$

上述公式中，fi_1 表示积极帖子的关注度，p_1 表示该帖子被判断为积极情绪的概率，fi_2 表示消极帖子的关注度，p_2 表示该帖子被判断为消极情绪的概率，$norm()$ 函数表示对数据进行零均值单位方差标准化，目的是消除不同指标之间量纲的影响。本文设计的投资者情绪指数规则如下：首先对每条帖子的关注度指数和概率值求积得到该帖的情感得分，其次对一天中所有帖子的情感得分求和得到该日的情感得分，然后对所有日期的情感得分进行标准化处理，最后与经过标准化处理的每日成交量相加，即可得到投资者情绪指数。

三、投资者情绪指数实证

（一）数据采集和清洗

本文选取用户最多且人气最高的证券论坛“东方财富网”中的“上证指数吧”作为数据源，

通过python编写的爬虫程序获取2018年7月16日至2019年7月15日的所有用户帖子约83.4万条样本数据，每条样本包括：标题、内容、作者、发帖时间、最后评论时间、阅读量和评论量。通过对数据进行初步观察，发现绝大多数帖子的标题和内容存在相同或者高度重复的现象，因此只保留帖子的内容部分来分析投资者情绪。进一步观察发现，该股吧帖子内容主要分为三种：股吧官方发布的信息帖、投资分析师的分析帖以及散户投资者的交流帖，三者的表现在字数上的区别在于：前两种帖子篇幅较长，通常在几百字甚至更多；而散户交流帖则偏短，一般为一到两个句子，总数不超过50字，为了剔除非散户帖子对研究结果的影响，本文将长度大于50字的帖子视作非散户发帖并且进行剔除。为了避免跨多天的交流帖对研究结果的影响，将发帖时间和最后评论时间不在同一天的样本剔除。由于股市逢节假日会关闭交易，而股吧每天都可以参与评论，为了统一二者的时间，需要删除发帖时间不在交易日的所有帖子。最后，对无效噪音帖进行筛选和剔除，这类帖子与股市内容无关，且大多以问句为主要形式，因此剔除所有带问号的帖子。经过上述数据清洗，此次研究得到了有效样本个数为696,287的样本库。

（二）Text-CNN 实例模型搭建

对样本数据进行预处理之后，从中随机选取一万条数据进行分类打标，作为模型的训练样本。本文将投资者情绪分为三种：积极、中性以及消极，以1表示积极、0表示中性、-1表示消极。三种情绪划分的标准为：将含有乐观情绪或者看涨情绪、觉得股市有利可图的帖子视作积极；将含有悲观情绪或者看跌情绪、觉得进一步投资会赔钱的帖子视作消极；将不带有任何情感偏向，只做简单陈述或者表达客观观点的帖子视作中性，选取部分帖子内容如表1所示。

表1 部分散户交流帖示例

编号	内容	分类	类型
1	中国股市牛起来！	1	表达乐观情绪
2	买买买，快加仓，大盘要拉红了。	1	表示有利可图
3	明天应该还会涨	1	表示看涨情绪
4	反复割韭菜行情，大家别买，买进去就跌。	-1	表示看跌情绪
5	给大家一个建议全部清仓，不然血本无归	-1	表示进一步投资会赔钱
6	清仓，再见！再好的大盘我也不会回来了	-1	表示悲观情绪
7	大盘早上还是突破 3000 点，下午接着回调	0	表达客观观点
8	看样子股指期货又要登台了	0	简单陈述

在进行模型训练之前，需要将样本数据处理成可输入模型的形式。首先利用python的第

三方库jieba分词对所有样本和训练集样本进行分词处理，在分词之前导入股市的专业术语和股民常用词，例如：利好、多头、癌股、大A等。同时，为了保证样本的语义完整，不进行停用词的操作。其次，利用第三方库keras的Tokenizer函数为分词之后的样本构建词汇表，得到词汇个数为1896的词汇表（部分词汇表内容见表2），并将样本中的词汇替换为词汇表中的对应的数字，从而将由自然语言表示的句子转换为由数字组成的句向量。同时，为了达到Text-CNN模型的输入要求，需要对句向量进行填充和截断处理，由于样本平均长度为27.4，因此本文设置的截断值为30，对长度超过30的句子进行截断，对不足30的句子用0进行填充处理。最后，还需要对样本数据集的分类指标进行one-hot编码，以便于模型的对种类的识别。经过上述步骤处理，得到了进行模型训练的训练集样本以及可以通过Text-CNN模型进行情感分类的所有样本数据。

表2 部分词汇表

词汇	Id	词汇	Id	词汇	Id	词汇	Id	词汇	Id
大盘	1	点	2	市场	3	明天	4	反弹	5
跌	6	指数	7	行情	8	涨	9	买	10
跳水	11	板块	12	a 股	13	机会	14	收盘	15
缺口	16	资金	17	股市	18	中国	19	业绩	20
韭菜	21	上证	22	抄底	23	趋势	24	股指	25

注：各词汇后面的数字为其在词汇表中对应的 id，id 由该词汇在样本库中所出现频次的顺序决定。

此次研究利用python的第三方库keras搭建Text-CNN模型。首先在词嵌入层中，将词向量的维数设置为300维，每个经过词嵌入处理后的句向量转化成一个30*300的二维张量。其次在卷积层中，设置窗口大小为2、3、4的三类卷积核各256个，分别对第一步得到的二维张量进行卷积，并使用‘Relu’函数进行激活，从而得到三个30*256的二维张量。然后在池化层中，对上述三个二维张量进行1维MaxPooling池化操作，得到三个1*256的二维张量。接着，对上述得到的三个二维张量进行拼接，得到1*768的二维张量，再通过“Flatten”层对其进行压缩，将其转换为一维向量。同时，在全连接层之前添加一个高斯噪音的drop层，随机对输入数据进行丢弃从而避免模型的过拟合。最后，将结果输入到全连接层，设置三种类型输出，通过“softmax”函数求出样本属于各种类别的概率值。后续只要选定概率值最大的类型作为样本的情感种类，即可得到样本的情感分类结果。本文构建的模型结构图如图2所示。经过多次调参，本文设计的Text-CNN模型的最优分类准确度为69.83%，涉及的超参数以及调参后的参数值如表3所示。

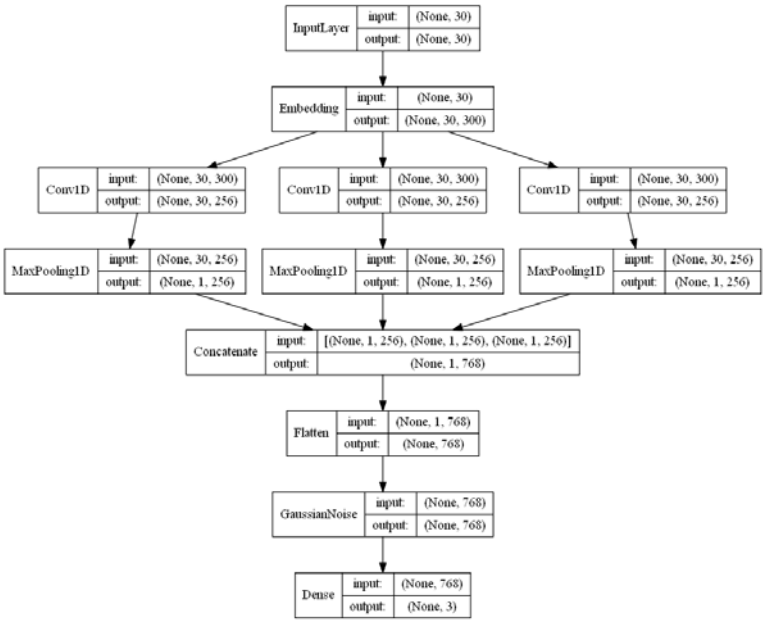


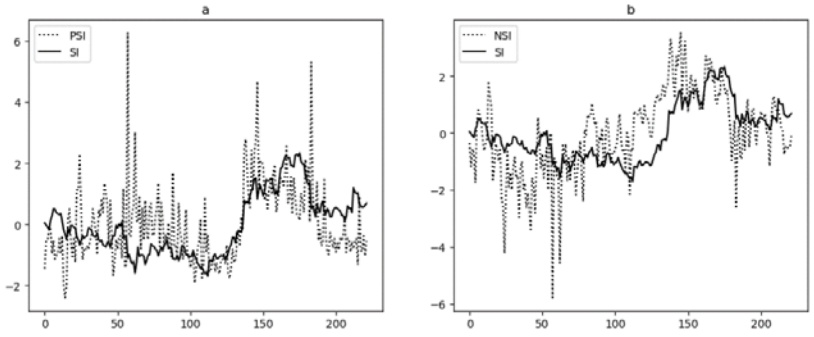
图2 Text-CNN实例结构图

表3 Text-CNN实例超参数及数值

参数	值	参数	值
batch_size	500	drop_rate	0.7
Epochs	20	卷积核个数	256

（三）投资者情绪指数构建

本文从“上证指数吧”获取2018年7月16日至2019年7月15日的散户投资者发帖数据，并且通过公式(2)和(3)构建了两种投资者情绪指数：积极投资者情绪指数(PSI)和消极投资者情绪指数(NSI)。为了研究投资者情绪指数与股指之间的关系，采集同期的上证指数(SI),对其进行零均值单位方差标准化后与投资者情绪指数进行对比研究，绘制投资者情绪指数与上证指数图如下所示：



（a）积极投资者情绪指数(PSI)和上证股指(SI)的对比（b）消极投资者情绪指数(NSI)和上证股指(SI)的对比

图 3 投资者情绪指数与上证指数对比图

图3中，a图为积极投资者情绪指数(PSI)和上证股指(SI)的对比图，其中虚线表示PSI指数，实线表示SI指数；b图为消极投资者情绪指数(NSI)和上证股指(SI)的对比图，其中虚线表示NSI指数，实线表示SI指数。由图3可以看出PSI、NSI和SI都有着较好的拟合。使用皮尔森相关系数检验投资者情绪指数和股指之间的线性关系，检验结果如表4所示。

表 4 情绪指数和上证指数间皮尔森相关关系

	皮尔森相关系数	P 值
PSI/SI	0.3143	1.7678335103990265e-06
NSI/SI	0.5779	3.477614100438587e-21

由表4可知，PSI与SI的皮尔森相关系数为0.3143，p值远小于0.01，同时NSI与SI的皮尔森相关系数为0.5779，p值同样远小于0.01，可以认为PSI、NSI都与SI有着正相关的线性关系。

四、实证检验和结果

（一）格兰杰因果检验

为了研究投资者情绪与股市价格之间是否存在统计意义上的因果关系，以及二者之间存在怎样的因果关系，本文对所构建投资者情绪指数和上证指数进行格兰杰因果关系检验。格兰杰因果关系检验需要遵循以下步骤：首先对这序列变量进行 ADF 单位根检验判断其平稳性，如果为平稳序列则可直接进行格兰杰因果检验；如果为非平稳序列，则需要对其进行协整检验，满足协整检验后可对原序列进行格兰杰因果检验[20]。单位根检验的原假设为：序列存在单位根，为非平稳序列，相关检验结果如表 5 所示：

表 5 原序列单位根检验结果

	ADF 检验统计量		显著性水平临界值		
序列变量	T 统计量	P 值	1%	5%	10%
PSI	-2.714607	0.071544	-3.460849	-2.874953	-2.573919
NSI	-2.145486	0.226621	-3.460849	-2.874953	-2.573919
SI	-1.529838	0.518702	-3.460707	-2.874891	-2.573885

由表 5 单位根检验结果可以看出，PSI、NSI 和 SI 的 p 值均大于 0.05，同时 NSI 和 SI 的 T 统计量均大于 10%的显著性水平临界值，PSI 的 T 统计量大于 5%的显著性水平临界值，因此三者均不能拒绝原假设，即三者都是非平稳序列。鉴于 PSI、NSI 和 SI 的原序列均为非平稳序列，考虑对其进行协整检验，由于协整检验要求参与检验的各序列是同阶单整，现对其一阶差分序列进行单位根检验，检验结果如表 6 所示：

表 6 一阶差分序列单位根检验结果

	ADF 检验统计量	显著性水平临界值

序列变量	T 统计量	P 值	1%	5%	10%
PSI 一阶差分序列	-12.380429	5.043518e-23	-3.460849	-2.874953	-2.573919
NSI 一阶差分序列	-12.587082	1.855767 e-23	-3.460849	-2.874953	-2.573919
SI 一阶差分序列	-6.872398	1.504002e-9	-3.461136	-2.875078	-2.573986

由表 6 单位根检验结果可得，PSI、NSI 和 SI 的 T 统计量均小于 1%的显著性水平临界值，且三者的 p 值远小于 0.05，因此在 1%的显著性水平下拒绝原假设，即 PSI、NSI 和 SI 的一阶差分序列为平稳的。由上述结果可知，PSI 和 SI、NSI 和 SI 均为同阶单整序列，可进行协整检验，检验结果如表 7 所示：

表 7 PSI、NSI 和 SI 的协整检验结果

	检验统计量		显著性水平临界值		
序列变量	检验值	P 值	1%	5%	10%
PSI 和 SI	-3.423122	0.039846	-3.946682	-3.363917	-3.063696
NSI 和 SI	-2.288546	0.379281	-3.946682	-3.363917	-3.063696

由表 7 协整检验结果可知，PSI 和 SI 的协整检验值为-0.3423122，小于 1%的显著性水平临界值，且 p 值为 0.039846 低于 0.05，因此在 1%的显著性水平下拒绝原假设，认定 PSI 与 SI 二者为协整关系，具备长期稳定的均衡关系，可进行格兰杰因果检验。同时，NSI 和 SI 的协整检验值为-2.288546，大于 10%的显著性水平临界值，且 p 值为 0.379281,明显大于 0.05，由此可判定二者不存在协整关系，无法进行格兰杰因果检验。对 PSI 和 SI 进行格兰杰因果关系检验，并且分别设置 1 到 4 阶滞后期，检验结果如表 8 所示：

表 8 格兰杰检验结果

原假设	滞后一阶		滞后二阶		滞后三阶		滞后四阶	
	F 统计量	P 值	F 统计量	P 值	F 统计量	P 值	F 统计量	P 值
PSI 不是 SI 的格兰杰原因	5.237	0.023	2.130	0.121	1.502	0.215	1.233	0.297
SI 不是 PSI 的格兰杰原因	5.692	0.000	10.650	0.000	5.361	0.001	4.130	0.003

由上表 1 到 4 阶滞后的格兰杰检验结果可知，“PSI 不是 SI 的格兰杰原因”的原假设在滞后一阶时的伴随概率为 0.0231 小于 0.05，因此在 5%的显著性水平下拒绝原假设，因此在滞后一阶时，PSI 是 SI 的格兰杰原因。但是在滞后 2 到 4 阶时，伴随概率值均大于 0.1，所以当滞后阶大于一阶时，PSI 不是 SI 的格兰杰原因。对于原假设“SI 不是 PSI 的格兰杰原因”，

在滞后 1 到 4 阶时的伴随概率值均小于 0.01，因此在 1%的显著性水平下拒绝原假设，即在滞后 1 到 4 阶，SI 都是 PSI 的格兰杰原因。由上述格兰杰检验结果可知，积极投资者情绪指数和上证指数之间存在双向的格兰杰因果关系，即积极投资者情绪和上证股指间存在互相的影响。反映到经济学意义为：投资者的积极情绪和股指存在相互影响关系，根据格兰杰因果检验，投资者的积极情绪会在一天的滞后期时表现到股指的变化上；同时股指因素会在 1 到 4 天的滞后期上反映到投资者的积极情绪的变化上。

（二）脉冲响应和方差分解分析

由格兰杰因果分析可知，投资者的积极情绪和股指之间存在着互相的影响关系，为了分析二者之间存在怎样的影响关系以及在互相影响中的作用有多大，本文通过构建向量误差修正模型(VECM)进行脉冲响应分析和方差分解分析，VECM 模型主要用于具有协整关系的非平稳时间序列的建模。利用 PSI 和 SI 构建 VECM 模型，并对其进行 F 矩阵的单位根检验，若单位根均小于 1 则说明模型是稳定的。由图 4 单位根检验结果可知，本文构建的 VECM 模型是稳定有效的。

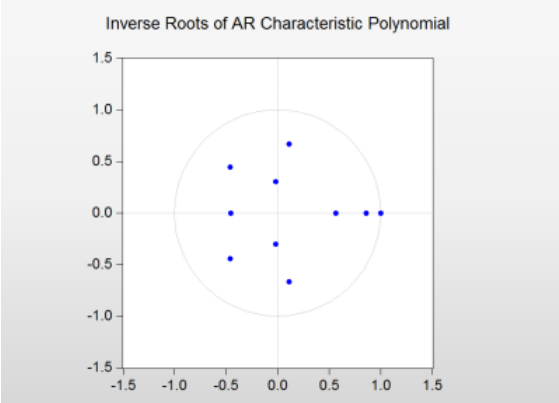


图 4 VECM 单位根检验结果

脉冲响应用于分析变量之间存在怎样的影响作用。在 VECM 模型的基础上得到 PSI 和 SI 的脉冲响应图如图 4(SI 对 PSI 脉冲的响应图)、图 5(SI 对 PSI 脉冲的响应图)所示。

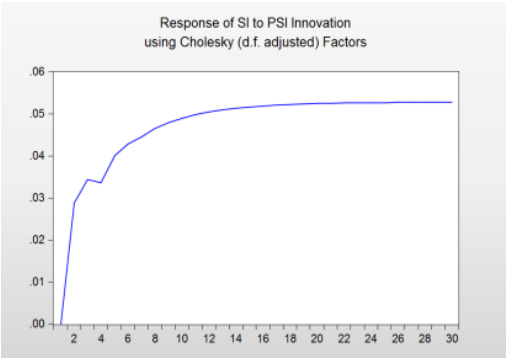


图 4 SI 对 PSI 脉冲的响应图

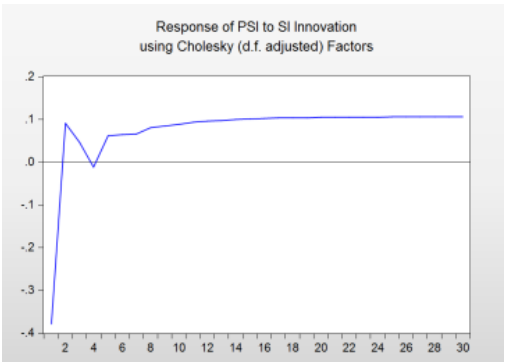


图 5 PSI 对 SI 脉冲的响应图

由图 4 可知，当受到投资者积极情绪的冲击时，上证指数从第 1 期开始上升在第 3 期和第 4 期中间产生一个小波动后，持续上升，在 17 期后趋于平稳。该脉冲结果表示，投资者的积极情绪对上证股指的正向冲击作用明显，且具有较长的持续时间。由图 5 可得，上证指数对投资者积极情绪在第 1 期时形成一个负向的冲击并逐渐减小降为 0，随后在第 2 期产生正向的冲击并达到最高点，然后再削减，在第四期降至 0 左右后再次于第 5 期升高并趋于稳定。由此可知，股指对积极情绪的影响在短期内会有波动，但总体呈现正向的冲击，最后形成持续的正向作用。综合上述结果可认为，投资者的积极情绪和股指之间的作用是相互且正向的，同时且这种正向作用将会长期存在。

方差分解的作用是解释变量在变化的过程中，自身和其他变量为其变化所做的贡献率为多少。通过 VECM 模型得到的方差分解图如图 6(SI 方差分解图)、图 7(PSI 方差分解图)所示。

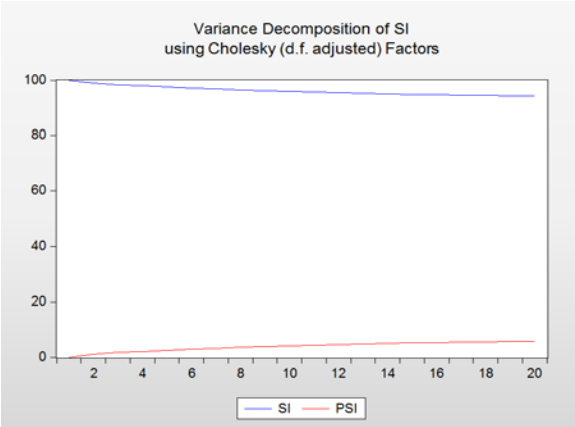


图 6 SI 方差分解图

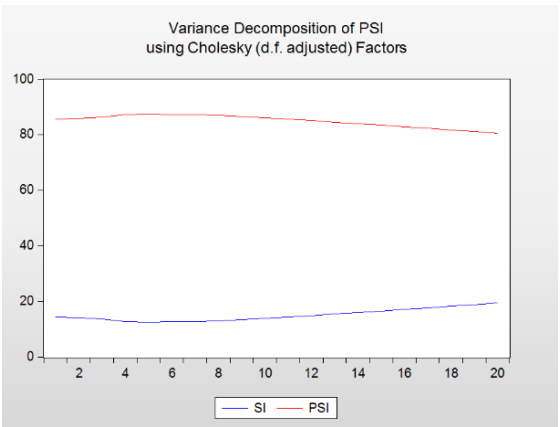


图 7 PSI 方差分解图

由图 6 可知，上证股指方差分解表明，其变动方差大部分原因来源于自身的，但自身变动可解释的部分呈现逐步下降的趋势，且积极投资者情绪指数的贡献在逐步增加，最终达到 6%左右。同样地，在图 7 中，积极投资者情绪指数变动方差很大部分由其自身所解释，股指贡献率在初期轻微下降后，呈现逐步上升的趋势，最后达到 20%左右。通过对股指和积极投资者情绪指数进行方差分解可得出这样的结论：尽管积极投资者情绪和股指对彼此发生变化的解释作用不是特别大，但随着时间推移都呈现出逐渐上升的趋势，说明在没有其他外界因素的干扰情况下，双方会对彼此施加一个长期累加的作用。反映到股票市场上可解释为：在没有利空消息披露和重大风险形成时，散户投资者的积极情绪和股指对彼此会施加一个长期的不断增强的促进作用，造成散户的投资热情不断升温同时股票指数的不断攀高的现象。

五、结论

论文将深度学习算法引入投资者情绪指数的研究。具体以论坛股吧为数据源收集散户投资者的发帖文本，通过建立 Text-CNN 模型对散户帖子内容所表达出的投资者情感进行分类，并在此基础上分别构建了积极投资者情绪指数和消极投资者情绪指数。研究结果表明，投资者情绪与股指存在正相关的线性关系。其中，乐观投资者情绪与股指之间存在长期且正向的相互影响作用，并且这种作用在不受其他因素干扰时呈现出逐渐增强的趋势。因此，在股票市场中，散户的积极情绪会促进股指的上涨，同时，上涨的股指会反过来促进散户对股票市场的积极态度，二者在这种相互影响中不断提高。论文的不足之处在于模型的情感分类效果仍存在很大的提升空间，同时由于提出的消极投资者情绪指数未通过协整检验，未能对散户投资者消极情绪与股指之间的关系进行深层次的研究，后续研究将着手弥补这些不足。

参考文献

[1] De Long，J.B.，Shleifer, A.，Summers.L.G.，Waldmann, R.J.，1990，“Noise Trader Risk in Financial Markets”，*Journal of Political Economy*，98(4)，pp. 703~738.

[2] Werner Antweiler，Murray Z Frank，2004，“Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards”，.*Journal of Finance*，59(3)，pp.1259~1294.

[3] Tetlock P C，2007，“Giving Content to Investor Sentiment: The Role of Media in the Stock Market”，*Journal of Finance*，62(3)，pp.1139~1168.

[4] 孟雪井、孟祥兰、胡杨洋：《基于文本挖掘和百度指数的投资者情绪指数研究》，《宏观经济研究》，2016 年第 1 期。

[5] 石善冲、朱颖楠、赵志刚：《基于微信文本挖掘的投资者情绪与股票市场表现》，《系统工程理论与实践》，2018 年第 38 期。

[6] 陈晓红、彭宛露、田美玉：《基于投资者情绪的股票价格及成交量预测研究》，《系统科学与数学》，2016 年第 36 期。

[7] 杜伟岸、吕佳玲：《社交媒体大数据、投资者情绪与 IPO 抑价》，《北京邮电大学学报(社会科学版)》，2018 年第 20 期。

[8] 董理、王中卿、熊德意：《基于文本信息的股票指数预测》，《北京大学学报(自然科学版)》 ,，2017 年第 2 期。

[9] Antweiler W，Frank M Z.，2004，“Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards”，*Journal of Finance*，59(3)，pp.1259~1294.

[10] 刘志明、刘鲁:《基于机器学习的中文微博情感分类实证研究》,《计算机工程与应用》，2012 年第 48 期。

[11] 徐军、丁宇新、王晓龙：《使用机器学习方法进行新闻的情感自动分类》，《中文信息学报》，2007 年第 6 期。

[12] Bengio. Y，Vincent P，Janvin C.，2003，“A neural probabilistic language model”，Journal of Machine Learning Research，3(6)，pp.1137~1155.

[13] Morin.F，Bengio.Y.， 2005，“Hierarchical Probabilistic Neural Network Language Mode” International Conference onArtificial Intelligence and Statistics.

[14] Goller C， Kuchler A.，1996，“Learning task-dependent distributedrepresentations by backpropagation through structure”[J].Neural Networks，1pp.:347~352.

[15] 梁军、柴玉梅、原慧斌等：《基于深度学习的微博情感分析》，《中文信息学报》，2014

年第 28 期。
[16] 李杰、李欢：《基于深度学习的短文本评论产品特征提取及情感分类研究》，《情报理论与实践》，2018 年第 41 期。
[17] Kim Y ，2014，“ Convolutional Neural Networks for Sentence Classification”. *Eprint Arxiv.*
[18] Zhang Y，Wallace B，2015，“A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification”， *Computer Science,*.
[19] 吴冲锋、吴文锋：《基于成交量的股价序列分析》，《系统工程理论方法应用》，2001 年第 1 期。
[20] 马若微、张娜：《我国股票市场投资者情绪 SENT 指数的构建——基于上证 A 股公司的面板数据》，《中央财经大学学报》，2015 年第 7 期。