时间序列

马晨轩 崔一帆

摘要：本文为探究是否能够利用股民投资情绪对股票走势进行预测，以信立泰（002294）为例，首先对其收益率数据建立 ARMA 模型；之后爬取了该股票 2019 年至 2020 年 2 月的网络股评，并利用每一天的情感评分计算股民投资情绪指数，将其作为外生变量代入 ARMA 模型，从而分析股票走势与投资情绪的关系；最后根据该模型进行短期预测，预测结果可为投资者做投资决策提供一定参考。

关键词：股坛评论，情感分析，股票趋势，ARIMA 模型，外生变量

Abstract: In order to explore whether investors' investment sentiment can be used to predict the stock trend, this paper first establishes an ARMA model for the yield data of XinLiTai(002294).Then this article crawled the online stock evaluation of the stock from 2019 to February 2020, calculated the investors' investment sentiment index by using the emotion score of each day, and substituted it into the ARMA model as an exogenous variable to analyze the relationship between stock trend and investment sentiment.Finally, the short-term prediction is made according to the model, and the prediction results can provide some references for investors to make investment decisions.

1 引言

随着我国经济社会和互联网技术的快速发展，互联网金融也拥有了良好的成长环境，不但有越来越多的个人以及机构投资者利用股票进行投资，也有越来越多的投资者倾向于在网络论坛上表达自己对股票的观点和预期。股票论坛的活跃度也日益上升，这些使结合投资者情绪的股票预测模型变得可行。

股票价格的波动和风险是通过时间显现的，同时大部分的股票收益率都是平稳序列，所以本文选择使用时间序列对股票收益率搭建模型。但单纯使用时间序列，及通过历史数据预测未来趋势并不能涵盖影响股票价格的所有信息，所以本文考虑加入外生变量提高模型预测的准确性.

影响股市走势的因素包括宏观经济因素、行业发展因素，公司内部因素以及股民情感因素，前两者因素短期内不会改变，公司内部因素一方面难以获取，另一方面难以数据化 [1]。同时股民情感因素又可以间接反映出其他三者数据。随着交易平台的发展，其社区和互动平台也得到了完善，股民所留的帖子和评论具有极高的时效性和个人情感，是分析股民投资情绪很好的材料。所以本文认为选取投资情绪作为外生变量是较好的选择。

于是本文选择了 XX 论坛 XX 股票 2019 年全年的评论进行情感分析。获得每一天的投资情绪指数后将其作为外生变量代入针对该股票建立的 ARMA 模型中，实现该股票的收益率以及风险预测，从而达到提高收益率以及规避风险的目的。

2 文献综述

2.1 情感分值计算及投资者情绪

国内外在投资者情绪对股票收益率影响的方面已经有了许多研究，如 Antweiler and Frank,2004; Fisher and Statman,2011; 段江娇, 刘红忠, 曾剑平,2017; 段江娇, 刘红忠, 曾剑平,2014; 周凌寒, 李波,2018 等等。Antweiler and Frank[2] 的研究表明股票评论对股票收益率有显著影响；Fisher and Statman[3] 的研究表明投资者情绪对股票收益有影响；段江娇等人的研究表明股票当日收益率受当日论坛情绪影响，且为显著正相关；周凌寒等人的研究表明融合了投资者情绪的股票行情预测模型准确度更高。

其中在“段江娇, 刘红忠, 曾剑平,2017”[4] 的研究中，研究者选择了东方财富网股吧论坛的帖子，使用文本处理技术提取帖子情绪，并引入到时间序列模型中，实现了股票收益率的预测。但是其文本处理部分使用的传统的基于语义规则情感分析模型，即利用特定的行业情感词典，基于句子语义规则计算文本的情感分值。因为在自然语言处理领域来说，全面且具有权威性的行业词典是稀缺资源，所以在使用基于语义规则的情感分析模型时，往往需要自己将一些领域专业词语加入到词典中，而自己加的词语难免会有较强的主观性及片面性。另外，基于语义规则的情感分析模型仅仅通过已经制定的规则计算情感分值，而不能很好的理解句子的深层含义，从而丢失了部分信息。而使用基于机器学习的情感分析模型就可以较好的解决这种问题。与基于语义规则的情感分析模型不同，基于机器学习的情感分析模型是从原始文本中提取文本特征，之后使用训练过的机器学习模型进行情感分值的计算或情感倾向的分类。这种方法能够很好地提取出文本的深度特征，以此提高对文本中情感表达倾向理解程度。

王美今, 孙建军 (2019) 根据“央视看盘”节目构造 BSI 指标，指出沪深两市中投资者情绪变化能显著影响收益，这表明投资者情绪是一个影响收益的系统性因子。在本文的研究中，我们使用 BSI 指标表示某只股票在 t 时期的投资者情绪指数，即看涨百分比。

2.2 时间序列

时间序列部分：时间序列是指将某种现象某一个统计指标在不同时间上的各个数值，按时间先后顺序排列而形成的序列。自 BOX 和 Jenkis 在 1927 年发现了对时间序列进行分析、预测以及对 ARIMA 模型识别、估计和诊断的系统方法，时间序列就被广泛地应用于股票的预测。例如，杨琦和曹显兵（2016）以对大众公用（600635）建立了 ARMA-GARCH 模型来预测股票价格，为投资者提供参考和指导。钟骐（2017）针对潍柴动力 180 个交易日的收盘价数据建立了 ARMA-GARCH 模型。又因近几年文本分析技术的愈发成熟，也有很多学者将文本信息纳入时间序列模型当中进行研究。王洪伟与张对（2015）对上证 300 中 30 只股票的网络股评进行情感打分，并将结果纳入时间序列构建了 ARMAX-GARCH 和 ARMA-GARCHX 模型，较好的预测了三十只股票的收益。同时发现将情感打分纳入 ARMA 模型中预测效果更好，并认为网络股评的影响滞后期大多数在 0 期到 1 期。本文在此基础上采用不同的文本分析基础并建立 ARMA-GARCH 模型对 XX 股票的收益率进行预测。

3 原理与方法

3.1 时间序列

3.1.1 $ARMA(p,q)$ 模型

一般的 $ARMA(p,q)$ 模型为：

$$r_t = \phi_0 + \phi_1 r_{t-1} \cdots + \phi_p r_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{1}$$

其中 r_t 为金融资产的在 t 时期的收益率,$\{\varepsilon\}$ 为独立误差项，是独立同分布零均值白噪声序列。p 为自回归模型的阶数 $\phi_0, \phi_1, \ldots, \phi_p$ 为自回归模型的待定系数; q 为滑动平均模型的阶数 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_q$ 为滑动平均模型的待定系数。满足 $ARMA(p,q)$ 模型的随机序列称为 $ARMA(p,q)$ 序列。AR 模型和 MA 模型分别是 ARMA 模型在 p 为 0 和 q 为 0 下的特殊情况。

3.1.2 $ARCH$ 效应以及 $GARCH(m,s)$ 模型

$ARCH$ 效应：$ARCH$ 效应，即自回归条件异方差性，是指不同时刻下时间序列的方差不平稳，同时其序列是不相关的，但是方差的平方或者是绝对值是相关的 [5]，也就是说时间序列的不同时间的方差之间是序列不相关但序列相依的。一般使用 $Ljung-Box$ 统计量 $Q(m)$ 对时间序列 a_t 的残差序列进行检验。

ARCH 模型：

$$a_t = \varepsilon_t \sigma_t;\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 \tag{2}$$

其中 ε_t 是零均值单位方差的独立同分布白噪声,σ_t 即序列的波动率 $\alpha_0 > 0, \alpha_i \geq 0$。ARCH 模型具有波动率聚集和扰动序列具有厚尾分布的优点，能较好地体现金融资产的特点。其中包含的基本思想有：扰动序列 $a_t = r_t - E(r_t|F_{t-1})$ 是前后不相关的, 但是并不前后独立;a_t 序列的相依性描述为 $Var(r_t|F_{t-1}) = Var(a_t|F_{t-1})$，可以使用 a_t^2 的滞后值的线性组合来表示。$GARCH(m,s)$ 模型：

$$r_t = \mu_t\ ;\ a_t = \varepsilon_t \sigma_t\ ;\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{3}$$

GARCH 模型是 ARCH 的推广模型，将的滞后值也纳入了公式中，不但保留了 ARCH 模型波动率聚集等特点，同时也减少了模型拟合时所需要的阶数。

3、具有外生变量的 $ARMA-GARCH$ 模型由于金融资产的时间序列既具有序列相关性，且具有尖峰厚尾和波动率聚集的特征，所以本文采取了 ARMA 与 GARCH 结合的模型。又因为在以往研究中证明在 ARMA 模型中加入外生变量的效果比在 GARCH 模型加入外生变量的效果要好，所以同时在均值方程，即 ARMA 模型中加入了外生变量，以探究投资者情绪对金融资产收益率的影响。结合方程：

$$\begin{cases} r_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \eta \gamma_t \ \ (\gamma_t \neq 0,\ \phi_i, \theta_i \neq 0) \\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \ \ (\alpha_i \neq 0,\ \beta_i \neq 0) \end{cases} \tag{4}$$

其中 γ_t 是 t 时期金融资产的投资情绪指数,η 表示 t 时期股票的投资情绪指数对 t 时期金融资产收益率的影响程度在公式 $(1), (2), (3), (4)$ 中出现的 r_t 均为股票在 t 时期的对数日收益率，即：

$$r_t = \ln(\frac{p_{t-1}}{p_t}) \tag{5}$$

其中表示上一期的股票收盘价格, 代表当期的股票收盘价格。

4 模型的构建和预测

4.1 投资者情绪分析

4.1.1 爬取

a) 股票的选择：

国内有许多数据量庞大的股票论坛，如新浪财经、天涯社区、雪球、东方财富等等。本文选择东方财富旗下的股吧进行数据收集，因为此论坛拥有划分明确的个股吧，而且活跃度很高。在东方财富的的股吧中，有多达一千多个个股吧，本文从被投资者广泛接受的“沪深 300”之中选择了一只个适于 ARCH 和 ARMA 模型的股票进行研究——信立泰（002294.SZ）。

b) 爬取、数据预览

本文使用 Python 编写的爬虫程序从“信立泰”的个股吧中爬取了从 2019 年 1 月 1 日至 2020 年 2 月 20 日共 15k 条网络评论。因为某些帖子本身不具有可交流性、发布时间论坛不活跃而导致帖子下评论数为零，所以在本研究中将这种帖子作为评论。如下：

日期	评论内容	内容来源
2019 年 2 月 12 日 18：19	明天估计又是老套路，高开几个点儿，阴跌，拉尾盘	评论
2019 年 2 月 22 日 23：19	此吧人气不足，是介入的好时机！	标题
2019 年 2 月 22 日 13：54	他们一直在跑路，就怕这样阴跌不断。	标题
2019 年 1 月 18 日 12：06	年后再说了，年前肯定回不了本	评论
2019 年 1 月 18 日 9：43	今天可以冲破重围拨云见日	评论

4.1.2 文本分析

1. 文本预处理因为节假日及周末时股票停牌，而在这些非交易日时股票论坛并不会禁止发帖，因此股票评论和股票收益率二者不能完全按照日期匹配。王洪伟和张对在 2015 年的研究表明前一天权重最大。基于以上原因，本文将非工作日的股票评论归并到其上一个工作日的评论之中。

2. 帖子评论情绪的量化

a) Python 库 SnowNlp 简介

SnowNLP 是一个中文的自然语言处理 Python 库，其主要功能有：中文分词、词性标注、情感分类、文本分类、提取关键词、提取摘要、分割句子、计算文本相似度等等。本文的研究主要使用了其中情感分类的功能，计算股票评论的情感分值及情绪类别。

b) 构造情感词典

因为 SnowNlp 的模型是在商品评论的数据集上训练得来的，所以当处理一些特定领域的文本时精准度可能会降低。另外，在自然语言处理领域，特定的行业词典的资源是极其稀缺的。基于以上因素，在计算情感分值之前，本文先构建了一个新的综合的情感词典，并将其加入到训练集中对模型进行微调，使其适用于股票文本的处理 [6]。

本文构造的新的情感词典包括两个部分，其一为几种常用的基础情感词典，其二为手动扩充的股票领域的专业词汇。整体结构如下：

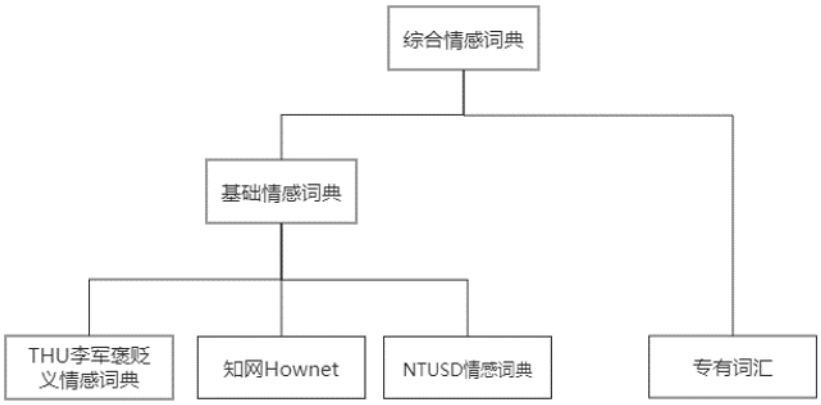

图 1: 情感词典结构

其中，清华大学李军褒贬义情感词典包含积极情感词汇和消极情感词汇各 5000 个左右，NTUSD 情感词典包含约 3000 个积极词汇以及 8000 个消极词汇，知网 Hownet 词典包括 800 个积极情感词汇、3000 个积极评价词汇、1200 个消极情感词汇以及 3100 个消极评价词汇。如下表所示：

积极词汇	消极词汇
内行、公平、可喜、巧妙、平衡、合理、有利、完整、一帆风顺、可以接受	暗下、暗中、悲伤、担心、担忧、下降、下陷、烦恼、无所谓、目光短浅、没意思

在此之外，本文扩充的股票专有词汇如下：

积极词汇	消极词汇
涨停、买空、盘坚、拉升、升仓、疯涨、杀进、利多、走高、翻红、牛市、赚钱、利好¬、反弹、长多、慢牛、见红	亏损、亏本、减产、破产、做空、卖空、长空、利空、吊空、清仓、平仓、下跌、抄底、走低、退市、套牢、砸盘、减持、盘软、回档、弱势股、概念爆发

c) 计算每一条股票评论的情感分值

做完以上准备工作之后，本文使用微调之后的 SnowNlp 模型对股票文本进行处理，计算出每一条评论的情感分值。情感分值大于零表示此评论的情绪为看涨，否则为看跌，如下表所示，标签列中的 0 代表看跌，1 代表看涨：

日期	评论内容	内容来源	情感分值	标签
2019 年 2 月 12 日 18：19	明天估计又是老套路，高开几个点儿，阴跌，拉尾盘	评论	-0.45768	0
2019 年 2 月 22 日 23：19	此吧人气不足，是介入的好时机！	标题	0.33137	1
2019 年 2 月 22 日 13：54	他们一直在跑路，就怕这样阴跌不断。	标题	-0.16088	0
2019 年 1 月 18 日 9：43	今天可以冲破重围拨云见日	评论	0.477243	1

4.1.3 投资者情绪指数的计算

投资者情绪指数是反应投资者意愿或预期的市场人气指标，对证券市场的运行和发展有很大的影响。本文参照央视看盘 BSI 指数（Bullish Sentiment Index）的方法计算投资者情绪 [7]，公式如下：

$$BSI_t = \frac{Bullish_t}{Bullish_t + Bearish_t} \tag{6}$$

其中，BSI_t 表示某一只股票在 t 时期的投资者情绪指数,$Bullish_t$ 表示此股票 t 时期的看涨帖子评论数量，$Bearish_t$ 表示此股票 t 时期的看跌帖子评论数量。BSI 越接近于 1，表示看涨情绪越强烈；越接近于 0，表示看跌情绪越强烈。

4.2 时间序列建模

4.2.1 数据选取

本文选取了 XX 股票作为研究对象，通过 wind 平台获取了 XX 股票从 2019 年 1 月 2 日到 2020 年 2 月 21 日所有交易日的收盘价格数据，并通过数据处理获得每一个交易日的对数收益率。以 2019 年 1 月 2 日到 2020 年 2 月 16 日的对数收益率数据为训练集，2020 年 2 月 17 日到 2 月 21 日的数据为测试集。

4.2.2 平稳性检验

在构建时间序列模型之前，必须保证该序列是弱平稳的。弱平稳性是指时间序列 Xt 的期望不随时间的改变而改变，且与的相关系数取决于时间间隔而非时间起始点。

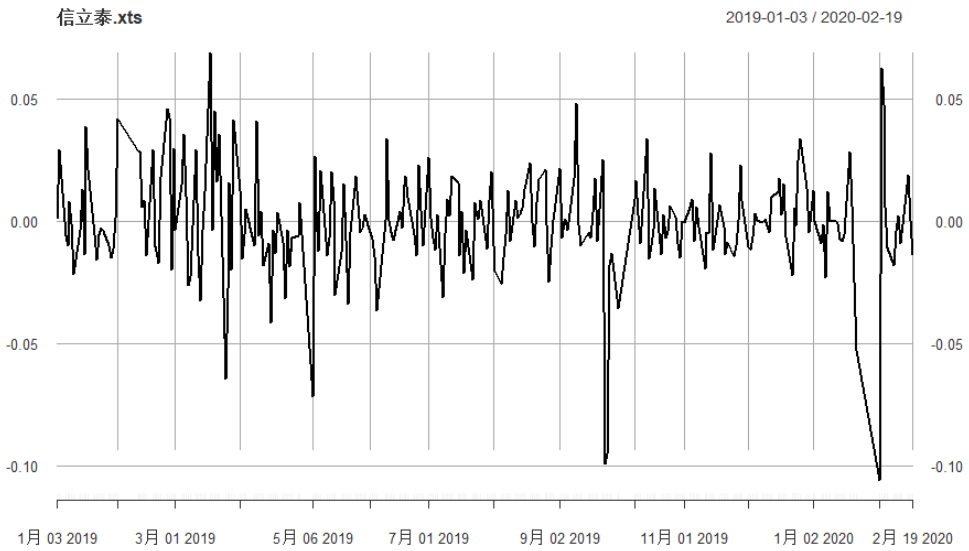

图 2: 信立泰时间序列图

图 1 为 XX 股票收益率序列的时间序列图，可以明显看出该收益率数据在 0 上下浮动，基本可以确定该收益率序列满足弱平稳的条件。

为了精确判断是否满足弱平稳，做出了序列的自相关以及偏自相关图，如图一所示。可以看出收益率序列几乎不存在相关性。对时间序列做单位根检验，得到 ADF 检验统计量为-6.8453，p 值为 0.01 显著，所以可以判断收益率序列具有弱平稳性。

4.2.3 建立均值方程

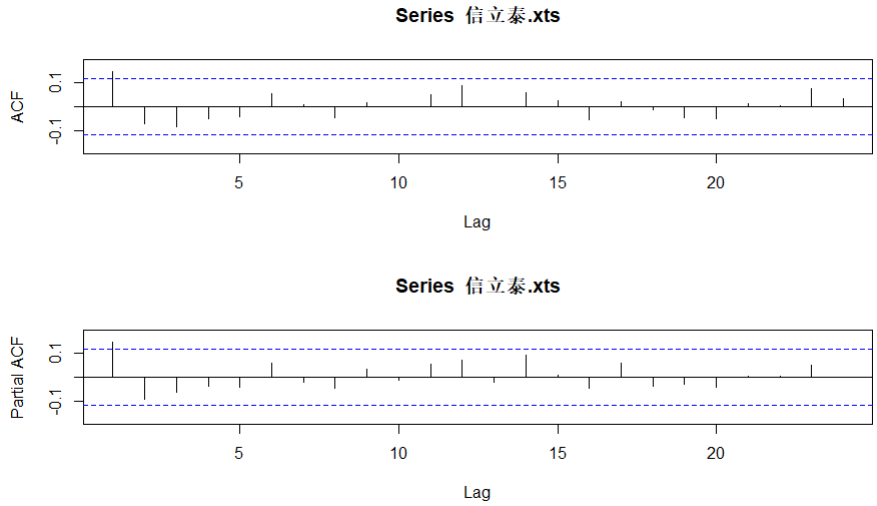

图 3: ACF 与 PACF 序列

从图一中的自相关分析图可以看出样本的自相关系数表现为截尾性，自滞后一阶开始不显著，则不考虑 AR 模型；从偏自相关分析图可以看出样本序列的偏自相关函数表现为拖尾性，在滞后 7 阶时仍表现显著，故利用 R 语言计算样本序列在加入外生变量时在不同模型下的 AIC 值。经过 AIC 信息准则计算，不同阶数的 AIC 值如下表所示：

表 1: 模型属性

模型	AIC 值	σ^2	对数似然函数值
$ARMA(1,0)$	-1322.71	0.000442	664.35
$ARMA(0,1)$	-1323.67	0.00441	664.83
$ARMA(1,1)$	-1321.99	0.00044	664.99

表 2: 各个参数值

(p,q)	ϕ_0	ϕ_1	θ_1
ARMA(1,0)	-0.0004	0.1457	-
ARMA(0,1)	-0.0004	-	0.1670
ARMA(1,1)	-0.0004	-0.1505	0.3115

由表一可以得知对该收益率序列建立 ARMA（1，1）的 AIC 值最小，且标准差与其他两种模型几乎相等，所以本文采用 ARMA（1，1）对收益率进行拟合。表二为使用最小二乘法拟合的各个模型的参数，可知该时间序列模型可表达为：

$$r_t = -0.0004 + \varepsilon + 0.1670\varepsilon_{t-1} \tag{7}$$

2> 对残差进行白噪声检验

在获得每一个参数值后需要对模型的残差进行白噪声检验，只有残差通过了白噪声检验才能够说明该时间序列模型是有效的，否则无法有效预测或需要对波动率进一步建模。本文使用 R 软件中自带的 Ljung-Box 检验对模型的残差进行白噪声检验，其原假设为数据为白噪声。

```
        Box-Pierce test

data:  XL.fit$residuals
X-squared = 0.013896, df = 1, p-value = 0.9062
```

图 4: 白噪声检验图

图三为模型残差的 Ljung-box 检验结果，P 值接近于 0，故可以认为模型残差中不存在尚未提取出来的信息，认定为白噪声。

进一步作出模型残差的正态 QQ 图和时序图 [8]。由图四正态 QQ 图可以得知残差序列不服从正态分布且具有厚尾现象。图五残差时序图可以得出较大的波动往往聚集在一起，于是收益率序列考虑是否存在 ARCH 效应。

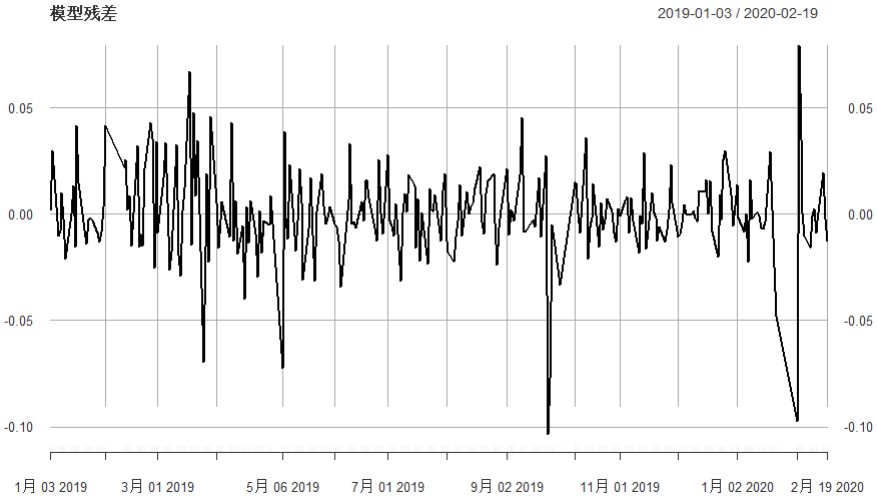

图 5: 残差时序图

4.2.4 波动率方程的建立

1> 检验是否具有 ARCH 效应考虑到残差序列可能存在相依性，于是通过 R 语言对模型的残差平方进行 Ljung 检验，检验的 P 值远小于 0.05，故可以拒绝原假设，认为时间序列具有条件异方差性。

2> 建立 GARCH 方程由于收益率序列具有条件异方差性，所以本文针对波动率建模。而是用 ARCH 模型会导致参数过多的问题，故采用 GARCH 模型进行建模。现在学术界尚未形成一套成熟的方法来确定 GARCH 模型的阶数且大部分模型都能够由 $GARCH(1,1)$ 解释，故本文在针对波动率建模时也采用 $GARCH(1,1)$。表三为 ARMA-GARCH 建模的各个参数大小。

该模型的方程表达式为：

$$\begin{cases} r_t = -0.0004 + \varepsilon + 0.1670\varepsilon_{t-1} \\ a_t = \varepsilon_t \sigma_t \\ \sigma_t^2 = 0.000227 + 0.4814 a_{t-1}^2 + 0.06291\sigma_{t-1}^2 \end{cases} \tag{8}$$

最后对模型的残差进行白噪声检验，其检验结果表明该模型的的残差为白噪声并说明不具有显著的在序列相关，所以所建立模型是有意义的。

4.2.5 含有投资情绪指数的 ARMA-GARCH 模型

重复实现上述四个步骤，同时在 ARMA 模型的拟合过程中加入外生变量的拟合，则模型公式可以表达为：

$$\begin{cases} r_t = 0.32 + 0.41 r_{t-1} + 0.12 r_{t-2} + \varepsilon - 0.72\varepsilon_{t-1} + 0.025\gamma_{t-1} \\ a_t = \varepsilon_t \sigma_t \\ \sigma_t^2 = 0.23 + 0.532 a_{t-1}^2 + 0.2341\sigma_{t-1}^2 \end{cases} \tag{9}$$

4.2.6 模型的预测

为了体现出包含投资情绪指数的时间序列模型能够更好的体现市场信息，更精确的预期收益率，分别运用包含投资情绪指数的和步包含投资情绪指数的时间序列模型对收益率序列进行一步与多步预测。首先根据建立好的不含有投资情绪指数行预测，之后将其还原为当日收盘价，并于实际值进行比对测算预测的精准程度。

日期	昨日收盘价	预测值	波动区间	实际收益率	预计收益率	标准差
2019 年 11 月 29 日	18.40	18.4528	0.0021	0.285%	0.8%	2.1%
2019 年 12 月 02 日	18.56	18.393	0.0021	-0.035%	-1.02%	2%

再根据含有投资情绪指数的时间序列模型进行预测:

日期	昨日收盘价	预测值	波动区间	实际收益率	预计收益率	标准差
2019 年 11 月 29 日	18.40	18.4532	0.0021	0.29%	0.8%	2.1%
2019 年 12 月 02 日	18.56	18.389	0.0021	-0.06%	-1.02%	2%

可以明显看出包含有投资情绪指数的时间序列模型能够更好的预测股票走势。但随着预测步长增加精确会随之减小，这是因为在多步测的过程中都是根据之前的预测值而不是实际值进行预测，但预测效果总体较好且多部预测的精准不会对股票的选择带来困难。

5 结论

本文将情感分析以及时间序列方法相结合，很好地挖掘了潜藏于评论中的投资情绪并将其作为外生变量输入时间序列模型。试验结果表明:

1> 股票论坛中的评论的确潜藏着有用的信息，且能够以投资情绪指数这一形式体现出来并加以运用。

2> 利用基于股票论坛评论计算得到的投资情绪指数确实能够使普通的时间序列模型得到更好的预测效果

6 参考文献

参考文献

[1] 王洪伟, 张对, 郑丽娟, and 陆颋, “网络股评对股市走势的影响: 基于文本情感分析的方法,” 情报学报, vol. 34, no. 11, pp. 1190–1202, 2015.

[2] W. Antweiler and M. Z. Frank, “Is all that talk just noise? the information content of internet stock message boards,” *The Journal of finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[3] K. L. Fisher and M. Statman, “Investor sentiment and stock returns,” *Financial Analysts Journal*, vol. 56, no. 2, pp. 16–23, 2000.

[4] 段江娇, 刘红忠, and 曾剑平, “中国股票网络论坛的信息含量分析,” 金融研究, vol. 10, pp. 182–196, 2017.

[5] 杨琦 and 曹显兵, “基于 arma-garch 模型的股票价格分析与预测,” 数学的实践与认识, vol. 46, no. 6, pp. 80–86, 2016.

[6] 周凌寒, “基于 lstm 和投资者情绪的股票行情预测研究,” Master’s thesis, 华中师范大学, 2018.

[7] 段江娇, 刘红忠, and 曾剑平, “投资者情绪指数, 分析师推荐指数与股指收益率的影响研究——基于我国东方财富网股吧论坛, 新浪网分析师个股评级数据,” 上海金融, vol. 11, pp. 60–64, 2014.

[8] 钟骐, “基于 arma—garch 模型的股票价格分析及预测,” 中国市场, no. 1, pp. 68–69, 2017.