基于 BVANet 的财经新闻情感分析

张 典[1]，王洁宁[1]，李昭颖[1]，刘润楠[2]，郑 文[1,3*]

(1. 太原理工大学大数据学院 太原 030060；2. 中国人民武装警察部队广东省总队参谋部 广州 510630；
3. 长治医学院健康大数据研究中心 山西 长治 046000)

【摘要】股票市场的预测一直以来是金融大数据分析领域一项难题，而财经新闻中包含的内在信息对市场表现有很大影响。提出了一种基于 BERT 的向量自回归融合网络 (BVANet)，该网络通过 BERT 将财经新闻情感量化，后结合市场表现联合构建金融时间序列向量自回归 (VAR) 模型，最终实现股票的预测。结果表明，与传统算法相比，BVANet 在提取新闻情绪信息和模型预测中取得了更好的效果，新闻的情绪对市场表现有预测作用。该研究可为自然语言处理在金融预测的应用提供实践参考。

关 键 词 深度学习； 财经新闻； 自然语言金融预测； 情感分析； 时间序列分析
中图分类号 TP183 **文献标志码** A doi:10.12178/1001-0548.2022058

A BERT-Based Vector Autoregressive Network for Sentiment Analysis of Financial News

ZHANG Dian[1], WANG Jiening[1], LI Zhaoying[1], LIU Runnan[2], and ZHENG Wen[1,3*]

(1. College of Data Science, Taiyuan University of Technology Taiyuan 030060;
2. Guangdong Province Corps General Staff Department, The Chinese Armed Police Force Guangzhou 510630;
3. Center for Healthy Big Data, Changzhi Medical College Changzhi Shanxi 046000)

Abstract Stock market forecasting is a difficult problem in the field of financial analysis. The intrinsic information contained in financial news has a great impact on the stock market performance. In this paper, we propose a BERT-based vector autoregressive network (BVANet), which quantifies financial news sentiment by BERT and then combines it with market performance to construct a financial time series vector autoregressive (VAR) model to achieve stock prediction eventually. The results show that BVANet has improved results in extracting news sentiment information and model prediction compared with traditional algorithms, and the sentiment of news has predictive effect on market performance. This study can provide a practical reference for the application of natural language processing in financial prediction.

Key words deep learning; financial news; natural language based financial forecasting; sentiment analysis; time-series analysis

有效市场假说 (efficient market hypothesis, EMH) 认为，股票市场价格受到所有可观察到的信息影响[1]。财经新闻涵盖了宏观经济走势、行业政策消息以及上市公司业绩状况等信息，这些新闻都会通过影响投资者情绪进而对金融市场产生作用。在最近十几年里，自然语言处理 (natural language processing, NLP) 发展迅速。文献 [2] 最早将分布式表征应用于单词，结合神经网络来训练语言模型。Mikolov 引入词向量 (word embedding) 技术[3]，实现了文本信息的向量表示，进而使得计算机可以更有效地理解和处理文本信息。Google 团队基于自注意力机制提出的 Transformer 结构[4] 以及 BERT 模型[5] 能更好地学习到句子中单词与单词之间的联系，从而能够结合上下文语境来提高挖掘信息的效果。NLP 中不断增强的文本表示能力使得计算机可以更准确地捕捉文本中的语义和情感。而如何将

收稿日期：2022－03－01；修回日期：2022－05－16
基金项目：国家自然科学基金 (11702289)；山西省关键核心技术和共性技术研发公关专项 (2020XXX013)
作者简介：张典 (1993－)，男，主要从事自然语言处理、金融大数据方面的研究.
*通信作者：郑文，E-mail：zhengwen@tyut.edu.cn

财经新闻中蕴含的信息高效准确地提取出来应用于金融市场是一项具有挑战性的研究课题。

在金融情绪分析研究中有一种基于情感词典提取的方法[6-8]，这种方法通过人工提取的方式构建金融领域的情感词表，通过一些加权方法得到文本的情绪指数，以此来进一步分析和金融市场的联系。这种方法只能捕捉到文本的表层特征，即词语的频次、重要程度等，无法获取文本的句法和语义特征。随着机器学习在金融领域的广泛应用，研究者提出了基于机器学习的方法来解决上述问题[9-12]。这种方法通过将文本视为词袋模型，采用贝叶斯分类[10]、逻辑回归和 SVM[12] 等算法训练文本分类器实现文本情绪提取进而分析金融市场。相比基于词典的方式，传统机器学习的方法可以捕捉到句法和语义层次的信息，提高情感表示的准确性。然而，研究表明深度学习在解决金融预测和分类问题时有更好的效果[13]。此外，当训练集较大时，深度学习的信息提取精度明显高于传统的机器学习方法[14]。因此，近年来研究者更倾向于采用深度学习的方法[15-19]。

已有研究表明，循环神经网络 (recurrent neural network, RNN) 相比卷积神经网络 (convolutional neural networks, CNN) 在捕捉上下文信息和建模复杂的时间特征方面更胜一筹[20]。其中文献 [20] 和文献 [15] 的工作比较有代表性。文献 [20] 提出了一种新的深度学习模型，通过段落向量将新闻文章转换成分布式表示，将长短期记忆网络 (long short-term memory, LSTM) 应用于金融时间序列预测中，并对多家公司开市价格的过去事件时间影响进行建模。文献 [15] 利用 LSTM 深度神经网络方法识别和提取金融新闻等文本的情感信息，构建自回归分布滞后模型和面板回归模型，从宏观市场以及微观股票资产两个层面实证揭示财经媒介信息所蕴含的情感对股票市场表现的关联影响。上述研究大多使用 RNN 或基于 RNN 的 LSTM 算法来提取文本信息，这些算法有如下缺点：1) 信息提取准确率不高，新闻文本较长，LSTM 在处理长文本时效果不好；2) 难以并行处理大规模的文本信息，而且效率低；3) 预测部分使用机器学习和深度学习的方法缺乏可解释性，不适用于金融分析。

为解决上述问题，本文构建了基于 BERT 的向量自回归融合网络 BVANet，该网络在情感信息提取过程中继承了 BERT 模型处理长文本的优势，利用自注意力机制提高了财经新闻情感信息提取的准确性和训练效率，之后通过量化情感信息并融合向量自回归 VAR 模型对市场表现进行时间序列分析。实验结果表明，相比于传统算法，BVANet 有更好的财经新闻情感分类效果，且在股票市场中有更好的可解释性和预测性。

1 数据和算法

1.1 数据收集

财经新闻训练集 (https://github.com/wwwxmu/Dataset-of-financial-news-sentiment-classification) 来自于雪球网发布的带有正负情绪的原始财经新闻。经过数据预处理后，得到了一个包含日期、公司、代码、标签、标题和文本的 17 149(正面 12 514；负面 4 635) 字段的数据集，用于 BERT 模型的第一阶段训练。

股票新闻数据集是 2020 年 1 月 1 日–2020 年 12 月 31 日从东方财富网上爬取到的个股新闻构建而成，用于时间序列分析，数据分布如表 1 所示。

表 1 股票新闻数据集分布

数据集	数据来源	样本	新闻数量/条
股票新闻数据集	东方财富网	贵州茅台	1 777
		山西汾酒	427
		五粮液	607
		洋河股份	384

1.2 BVANet 原理

1.2.1 算法流程

图 1 为 BVANet 的整体算法流程，总体上分为两个阶段。第一阶段是财经新闻情感量化阶段，首先使用自监督的训练方法对维基百科中文预料 进行训练，得到预训练词向量表征。为了赋予模型金融领域特征，采用有监督训练方法通过财经新闻训练集对预训练模型进行微调，得到财经新闻情感量化模型，用于股票新闻数据集中财经新闻的情感量化。第二阶段是时间序列分析阶段，为了研究新闻情感和股价之间的关系，通过数据处理层将量化的新闻情感时间序列化，与股价时间序列相对应。金融领域研究者通常需要模型具有科学依据和解释性，所以在 BAVNet 建模和预测部分采用了建立在统计学基础上具有可解释性的向量自回归 (VAR) 模型。

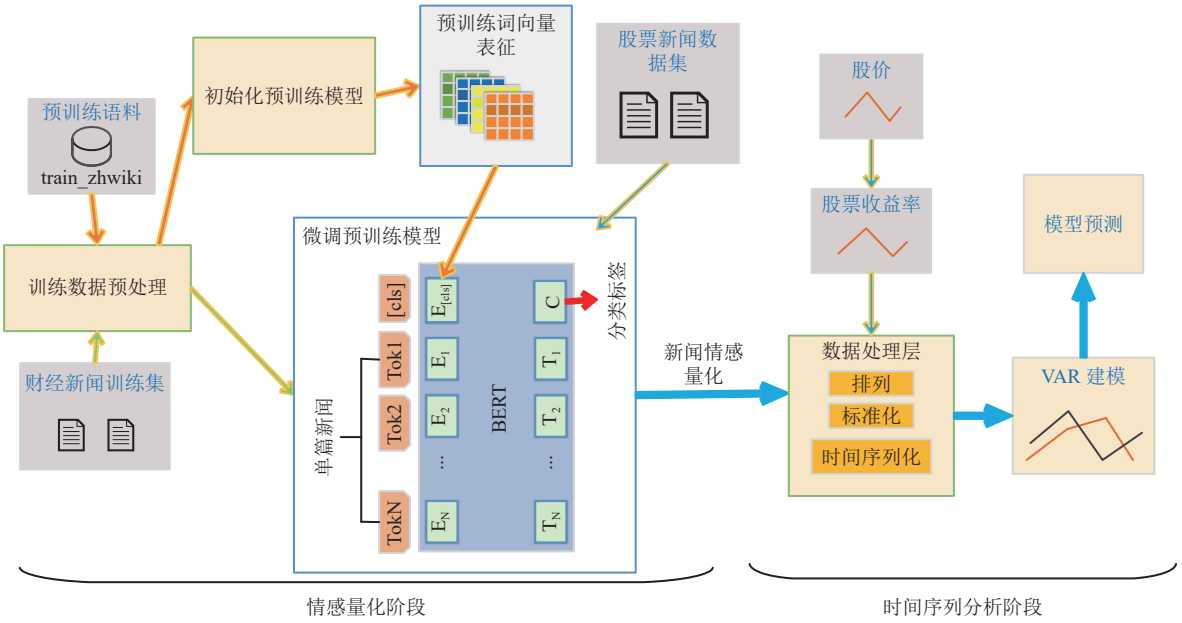

图 1　BAVNet 模型图

1.2.2　自注意力机制编码

自注意力机制用于将新闻编码为矩阵的过程中，如图 2 所示。输入的$\boldsymbol{X}\in\mathbb{R}^{(\mathrm{BS}\times\mathrm{SL})}$是一个自然语言序列，BS 表示每一次输入到模型中训练新闻的数量，SL 表示序列的长度。经过初始化词向量之后将句子序列中每个字转化成向量表示，得到输入新闻$\boldsymbol{X}$的字向量矩阵$\boldsymbol{X}_{\mathrm{emb}}\in\mathbb{R}^{\mathrm{BS}\times\mathrm{SL}\times\mathrm{dim}}$，dim 表示字向量维度。

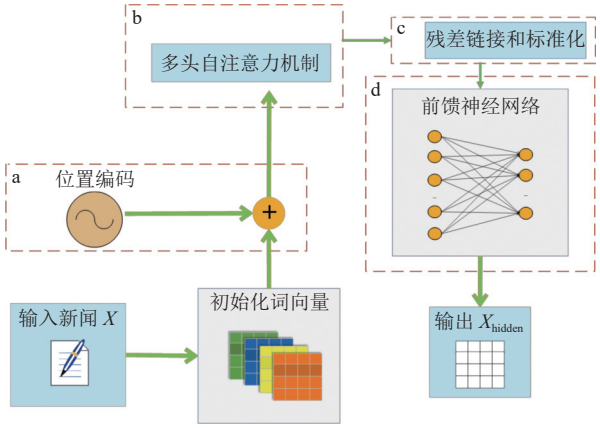

图 2　编码过程

图 2 中 a 框表示位置编码，用于捕捉序列位置信息[4]。加入位置编码可以让模型更好地理解较长的新闻文本。b 框是多头自注意力机制，为了学习多重含义的表达，初始化 3 个权重矩阵$\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V\in\mathbb{R}^{(\mathrm{dim}\times\mathrm{dim})}$；对$\boldsymbol{X}_{\mathrm{emb}}$做线性映射，得到 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 这 3 个矩阵：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{X}_{\mathrm{emb}}\boldsymbol{W}_Q\\ \boldsymbol{K}=\boldsymbol{X}_{\mathrm{emb}}\boldsymbol{W}_K\\ \boldsymbol{V}=\boldsymbol{X}_{\mathrm{emb}}\boldsymbol{W}_V\end{cases}\tag{1}$$

多头注意力机制首先将字向量维度平均分割成 h 份，h 是头的数量，多头机制是为了刻画一个字多个方面的含义。引入多头后 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 的维度为$\mathbb{R}^{\mathrm{BS}\times\boldsymbol{h}\times\mathrm{SL}\times\mathrm{dim}/h}$，注意力机制如式 (2)。先求出注意力矩阵$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$，注意力矩阵的第 i 行包含第 i 个字和所有字的相关性信息，是为了把注意力矩阵变成标准正态分布，以便反向传播梯度，softmax 归一化使每个字跟其他字的注意力权重的和为 1，得到注意力权重概率分布[4]。之后使用注意力矩阵给$\boldsymbol{V}$ 矩阵加权从而使每个字向量都含有当前句子内所有向量的信息。

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\tag{2}$$

图 2 中 c 框是残差连接和标准化，经过注意力矩阵加权后与 $\boldsymbol{X}_{\mathrm{emb}}$元素相加得到残差连接，训练时可以使梯度直接走捷径反传到初始层。标准化是把神经网络中隐藏层归一为标准正态分布，以起到加快训练速度、加速收敛的作用。d 框是前馈神经网络，利用两层线性映射并用激活函数激活得到新闻 X 的隐藏层编码矩阵$\boldsymbol{X}_{\mathrm{hidden}}$。本文将输入的自然语言序列 X 编码成为$\boldsymbol{X}_{\mathrm{hidden}}$。

1.2.3　BERT 模型情感分类

BERT 以遮蔽语言模型 (masked language model) 和下一句预测 (next sentence prediction, NSP) 两种

训练方式来建立语言模型[5]。NSP 是为了让 BERT 更好地学习句与句之间的关系，让模型预测句子 B 是否为句子 A 的下一句。进行情感分类任务时会使用 NSP 这种语言模型，如图 3 所示。BERT 模型在文本前插入一个 [CLS] 符号，并将该符号对应的输出向量作为整篇文本的语义表示，用于文本分类。与文本中已有的其他字相比，这个无明显语义信息的符号会更“公平”地融合文本中各个字的语义信息。从 $\boldsymbol{X}_{\text{hidden}}$ 中得到 [CLS] 对应表征 $\mathbf{cls\_}\boldsymbol{v}$，之后通过式 (3) 得到文本情感分类的推断 $\hat{y}$。

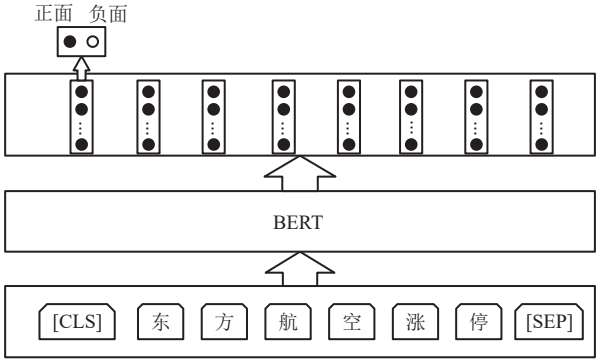

图 3 情感分类任务

$$\hat{y} = \text{sigmoid}(\boldsymbol{W} * \mathbf{cls\_}\boldsymbol{v}) \tag{3}$$

1.2.4 情感值量化和数据处理

训练数据是将财经新闻训练集中字段为“正文”和“标签”的数据取出，并进行随机打乱。按照 7:3 比例分成训练集和测试集，数据为正文内容和情感正负面，积极情感为 1，消极为 0。虽然没有任何中性情感的数据集，但在式 (3) 中，经过 sigmoid 函数后的值代表了其正面情绪程度。因此，当这个值接近分类阈值时可以被认为是偏中性的情绪。基于式 (3) 得到新闻 d 是正向新闻的预测概率 $\hat{y}_d \in (0,1)$，由 $\hat{y}_d$ 计算 d 的情感值 $\hat{S}_d$，如式 (4) 所示。当 $S_d > 0$，情感为积极 (正向)，$S_d < 0$，情感为消极 (负向)。

$$S_d = (-1,1) \times \hat{y}_d \tag{4}$$

在数据处理层部分，股票市场表现时间序列是通过 python 中财经数据接口包 Tushare 获取，企业的每个交易日的收盘价用 p_t 表示。与 p_t 相对应，将对应企业的所有新闻情感量化后按照时间顺序排列。由于交易日和新闻日期并非一一对应，本文将两个交易日之间所发布的新闻全都算前一个交易日时间片中。之后计算时间片 t 内全部 N_t 条新闻的总体情感值 S_t，如式 (5) 所示。依据时间序列计算累计情感值 S_T，如式 (6) 所示。以上过程为数据处理层的排列和时间序列化过程。由于 p_t 和 S_T 在数量级上差距过大，对其进行均值标准化，得到标准化时间序列 $p_{t,N}$ 和 $S_{T,N}$。

$$S_t = \frac{1}{N_t}\left(\sum_{d=1}^{N_t} S_d\right) \tag{5}$$

$$S_T = \sum_{i=1}^{T}\left(\frac{1}{N_i}\left(\sum_{d=1}^{N_i} S_d\right)\right) \tag{6}$$

1.2.5 时间序列分析

上述过程得到了时间序列数据，之后按照计量经济学相关理论进行时间序列分析，分析过程如图 4 所示。

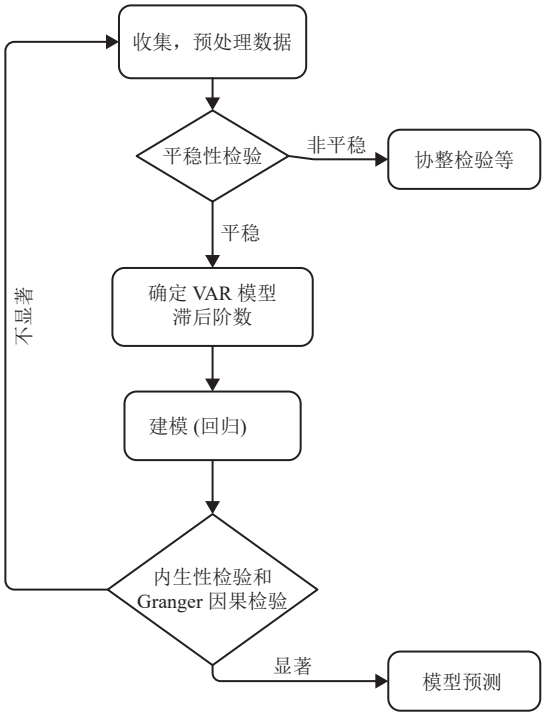

图 4 时间序列分析步骤

在计量经济学中 VAR 模型常用于预测相互关系的时间序列[21]。在 VAR 模型中，所有变量都由自身的滞后项和其他内生变量的滞后项及随机误差进行解释。设两个时间序列为 $\{y_{1t}, y_{2t}\}$，分别作为两个回归方程的被解释变量，解释变量为两变量的 p 阶滞后值，构成二元 VAR(p) 系统。

$$\begin{cases} y_{1t} = \beta_{10} + \beta_{11} y_{1,t-1} + \cdots + \beta_{1p} y_{1,t-p} + \\ \qquad \gamma_{11} y_{2,t-1} + \cdots + \gamma_{1p} y_{2,t-p} + \varepsilon_{1t} \\ y_{2t} = \beta_{20} + \beta_{21} y_{1,t-1} + \cdots + \beta_{2p} y_{1,t-p} + \\ \qquad \gamma_{21} y_{2,t-1} + \cdots + \gamma_{2p} y_{2,t-p} + \varepsilon_{2t} \end{cases} \tag{7}$$

内生性检验是检验 VAR 模型中解释变量对被解释变量是否有显著性影响。Granger 因果检验是检验某个变量的滞后值 (过去的信息) 对被解释变量的信息是否有预测能力。Granger 因果检验得出的因果关系不是实际经济活动中的因果关系。Granger 因果检验假设了有关 y 和 x 每一变量的预测信息全部包含在这些变量的时间序列之中。检验要求估计式 (8) 和式 (9) 的回归。如果拒绝原假设，则 x 是y_t的 Granger 因果关系，即 x 对y_t有预测能力，否则，x 对y_t没有预测能力。

$$y_t = \sum_{i=1}^{q} \alpha_i x_{t-i} + \sum_{j=1}^{q} \beta_j y_{t-j} + u_{1t} \tag{8}$$

$$\begin{cases} \text{原假设} H_0 : \alpha_1 = \alpha_2 = \cdots = 0 \\ \text{备择假设} H_1 : \text{至少有一个} \alpha_i \neq 0 \end{cases} \tag{9}$$

2　实　验

2.1　评价指标

AUC (area under roc curve) 被定义为受试者工作特征曲线 (receiver operating characteristic curve，ROC) 下的面积。AUC 是衡量二分类模型优劣的一种评价指标，表示预测的正例排在负例前面的概率[22]，其值越接近 1 表示模型的分类效果越好。

准确率 (accuracy) 指的是对于给定的数据集，模型正确分类的样本数与总样本数之比。它可以直接反应模型预测的准确程度，其值越接近 1 表示模型分类效果越好。

2.2　财经新闻情感提取

本次研究所用 BERT 模型基于深度学习框架 PyTorch 实现，训练 GPU 为 NVIDIA Tesla P4 8 GB 显卡，模型超参数设置如表 2 所示。

表 2　训练模型超参数设置

参数	值
输入字向量维度	384
Transformer层数	6
Multi-head个数	12
优化器	Adam
Batch Size	1
Dropout	0.4
学习率	1×10^{-6}

图 5 为 BVANet 模型在微调后训练集和测试集的 AUC、accuracy 指标在训练后的表现。训练集和测试集 AUC 收敛于 0.98，accuracy 均收敛于 0.95。结果表明训练的模型在分类效果上表现良好，训练集和测试集结果相近说明模型没有过拟合，泛化能力较好。

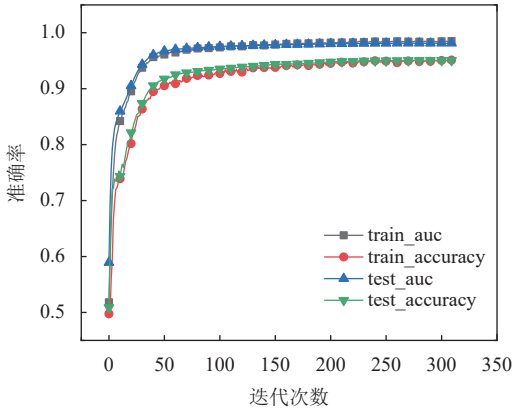

图 5　模型训练图

财经新闻情感分类结果与 SVM、BP、CNN、LSTM 算法对比结果如表 3 所示。可以看出，除了 Recall 外，其余指标使用 BVANet 模型都达到了更好的效果。

表 3　不同情感分析模型评估

算法	Accuracy	Precision	Recall	F1-score
SVM	0.588 2	0.592 0	09674	0.735 7
BP	0.590 2	0.589 6	**0.989 6**	0.745 0
CNN	0.753 5	0.814 5	0.755 8	0.782 4
LSTM	0.824 4	0.810 8	0.893 6	0.851 0
BVANet	**0.950 8**	**0.951 3**	0.952 1	**0.951 7**

2.3　新闻情感值和股票价格相关性分析

将 2020 年股票新闻数据集中的贵州茅台、五粮液、山西汾酒和洋河股份 4 只股票的所有新闻数据输入到训练好的模型中，量化每篇新闻情感值，并得到标准化后的股价历史数据时间序列 $p_{t,N}$ 和累计情感值时间序列$S_{T,N}$，4 只股票的累计情感值和收盘价时间序列可视化如图 6 所示。其中五粮液在四月份有明显情感和股价相反的趋势，经过溯源发现这段时间的新闻报导重点在于五粮液集团的负面报道，模型根据语义理解为负面情感，但实际是一种利好，因此出现了偏差。二者的 Pearson 相关性如表 4 所示，从表中可以看出二者在 0.05 的水平下相关性显著。

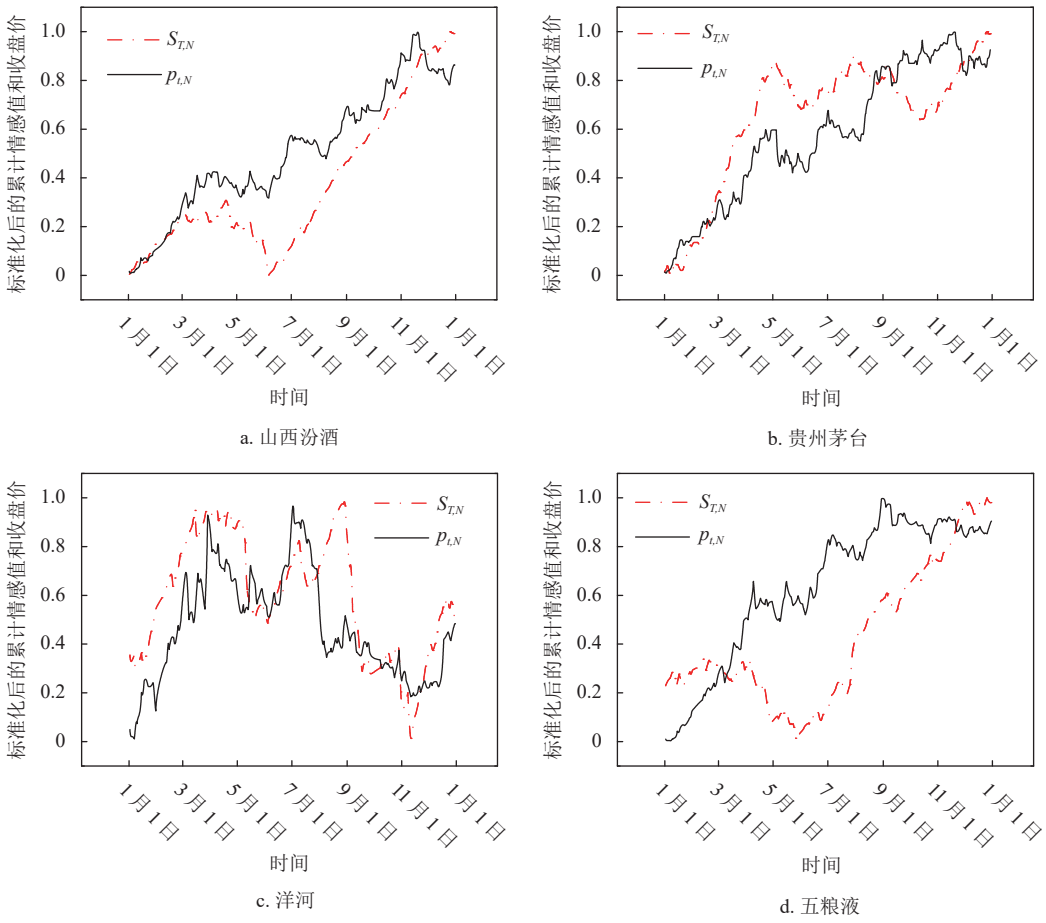

图 6 $S_{T,N}$和$p_{t,N}$的时间序列可视化

表 4 Pearson 相关性

个股	Pearson相关性
贵州茅台	0.793 792**
五粮液	0.565 619**
山西汾酒	0.880 881**
洋河	0.663 530**

**表示显著性水平为0.05

2.4 时间序列分析

上述结果表明新闻情感累计值和每日收盘价的相关性，以下通过时间序列分析两个变量之间内在的影响关系。

2.4.1 VAR 建模

以山西汾酒为例，引入对数收益率[7]：

$$R_t = \log(p_t | p_{t-1}) \tag{10}$$

式中，R_t是 t 时刻的对数收益率；p_t 表示当日收盘价；p_{t-1} 表示前一日收盘价。同样将其标准化后得到标准化后的对数收益率$R_{t,N}$，将累计情感值做一阶差分，得到情感变化时间序列$S_{T,N}{}^1$。以上两个时间序列在经过平稳性检验后均为平稳序列。

VAR 模型依据 AIC、SC 等信息量取最小准则来确定阶数，如表 5 所示，当滞后项 Lag=1 时达到最小。

表 5 信息准则

Lag	LR	AIC	SC	HQ
0	NA	−6.622 052	−6.592 961	−6.610 329
1	**25.148 49**	**−6.695 141**	**−6.607 866**	**−6.659 971**
2	2.527 013	−6.672 467	−6.527 009	−6.613 851
3	4.607 531	−6.658 855	−6.455 212	−6.576 792
4	1.972 506	−6.633 958	−6.372 132	−6.528 449
5	2.497 880	−6.611 441	−6.291 431	−6.482 486

确定对收益率和情感值变化建立 VAR (1) 模型。建模结果为：

$$\begin{cases} S_{T,N}{}^1 = 0.238 S_{T,N}{}^1(-1) + 1.97\times \\ \qquad R_{t,N}(-1) + 0.0429 \\ R_{t,N} = 0.00791 S_{T,N}{}^1(-1) - 0.118\times \\ \qquad R_{t,N}(-1) + 0.00144 \end{cases} \tag{11}$$

表 6 为该模型内生性检验，从表中可以看出显

著性水平为 0.05 以下认为解释变量$S_{T,N}^1$对被解释变量$R_{t,N}$有显著的影响。以上结果说明，财经新闻情感变化和股票收益率的变化是互相影响的，而情感变化对收益率变化的影响更为显著。

Granger 因果检验结果如表 7 所示。可以看出：1) 在滞后期为 1, 2 期情况下，分别以 5% 和 10% 的显著性水平检验结果拒绝了$S_{T,N}^1$ 不是 $R_{t,N}$ 的 Granger 原因的原假设，即新闻情感的波动在一定时间段内先行于股票收益的波动。而在 3, 4, 5, 6 期没有拒绝该假设，则说明新闻情感波动在短期内对个股收益率具有一定的预测作用，而在长期内则不存在影响；2) 滞后期为 1 期时，检验结果在 10% 的显著性水平拒绝了$R_{t,N}$不是$S_{T,N}^1$的 Granger 原因的原假设，而 2, 3, 4, 5, 6 期没有拒绝。说明部分情况下股价的变动将迅速反映在情绪波动上。

表 6　内生性检验

解释变量	被解释变量	滞后阶数	P值
$R_{t,N}$	$S_{T,N}^1$	1	0.071 6
$S_{T,N}^1$	$R_{t,N}$	1	0.028 4

表 7　Granger 因果检验

原假设H_0	参数	滞后期				
		1	2	3	4	5
$S_{T,N}^1$不是$R_{t,N}$的Granger原因	F统计量	4.80**	2.57*	1.68	1.55	1.71
	P值	0.029 3	0.079	0.171	0.189	0.134
$R_{t,N}$不是$S_{T,N}^1$的Granger原因	F统计量	3.25*	1.45	1.68	1.19	0.973
	P值	0.072 9	0.238	0.173	0.315	0.435

**表示显著性水平为0.05，*表示显著性水平为0.1

2.4.2　股价预测

根据 VAR(1) 的建模结果以及式 (10)，预测山西汾酒在 2021 年前 16 个交易日的股价，并与只考虑$S_{T,N}^1$和收盘价作为特征的基于 RBF 核的 SVR 算法和 LSTM 算法进行比较。结果如表 8 所示，表明在只用新闻情感值和股票价格作为预测特征的情况下，基于向量自回归模型的 BVANet 优于 SVR 和 LSTM。

表 8　山西汾酒股价预测结果

算法	MSE	RMSE	MAE	MAPE/%	SMAPE/%
SVR	5.051 3	2.247 5	1.961 6	2.12	2.13
LSTM	4.428 3	2.104 3	1.752 7	1.93	1.92
BVANet	**1.017 3**	**1.008 6**	**0.829 4**	**0.90**	**0.90**

3　结 束 语

本文提出的 BVANet 是一种基于 BERT 向量自回归融合网络的股票预测方法，本文使用其分析了个股 2020 年全年新闻情感与市场表现的关系。结果表明 BVANet 在财经新闻的情感量化和股票预测中均优于其他方法，财经新闻情感信息对股票市场表现具有预测作用。BVANet 完整地进行了模型的训练、模型评估、时间序列分析和股票预测的全过程，对自然语言处理在金融领域的应用提供了实践参考。

未来研究或可从以下 2 个方面深入展开：

1) 本研究在训练模型时选取的细粒度为字向量，在中文中一个词语可能比字更能表达清楚，接下来可以利用十分完备的分词工具对文本进行分词，选取细粒度为词构建词向量。进一步提升对文章的向量表达效果，提高情感表达的精确性，更好地把握情感。

2) 本研究在进行时序分析时只用了财经新闻的情感特征和市场表现特征，然而实际的市场远比这要复杂的多，接下来可以加入社交媒体信息、股吧评论信息等，更全面地探究不同类型信息以及它们对市场的不同影响。

参　考　文　献

[1] FAMA E F. The behavior of stock-market prices[M]. Chicago: University of Chicago, 1965.

[2] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[EB/OL]. [2021-10-9]. https://jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.

[3] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2021-10-11]. https://arxiv.org/pdf/1301.3781.pdf.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2021-11-11]. https://arxiv.org/pdf/1706.03762.pdf.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-Training of deep bidirectional transformers for language understanding[EB/OL]. [2021-11-18]. https://arxiv.org/pdf/1810.04805.pdf.

[6] 顾文涛, 王儒, 郑肃豪, 等. 金融市场收益率方向预测模型研究—基于文本大数据方法[J]. 统计研究, 2020, 37(11): 68-79.
GU W T, WANG R, ZHENG S H, et al. Research on the prediction model of the direction of financial market returns: Based on text big data method[J]. Statistical Research, 2020, 37(11): 68-79.

[7] 孟志青, 郑国杰, 赵韵雯. 网络投资者情绪与股票市场价格关系研究——基于文本挖掘技术分析[J]. 价格理论与实践, 2018(8): 127-130.

MENG Z Q, ZHENG G J, ZHAO Y W. The Research on the relationship between network investor emotion and stock market price: Empirical analysis based on text mining technology[J]. Price Theory & Practice, 2018(8): 127-130.

[8] 姚加权, 冯绪, 王赞钧, 等. 语调、情绪及市场影响: 基于金融情绪词典[J]. 管理科学学报, 2021, 24(5): 26-46.
YAO J Q, FENG X, WANG Z J, et al. Tone, sentiment and market impacts: The construction of Chinese sentiment dictionary in finance[J]. Journal of Management Sciences in China, 2021, 24(5): 26-46.

[9] 冉杨帆, 蒋洪迅. 基于 BPNN 和 SVR 的股票价格预测研究[J]. 山西大学学报 (自然科学版), 2018, 41(1): 1-14.
RAN Y F, JIANG H X. Stock prices prediction based on back propagation neural network and support vector regression[J]. Journal of Shanxi University A (Natural Science Edition), 2018, 41(1): 1-14.

[10] GIDOFALVI G, ELKAN C. Using news articles to predict stock price movements[EB/OL]. [2021-12-18]. https://www.researchgate.net/profile/Gyozo-idofalvi/publication/228892903_Using_news_articles_to_predict_stock_price_movements/links/54f58e690cf2ba6150668a52/Using-news-articles-to-predict-stock-price-movements.pdf.

[11] IZUMI K, GOTO T, MATSUI T. Trading tests of long-term market forecast by text mining[C]//The 10th IEEE International Conference on Data Mining Workshops. Sydney: IEEE, 2011: 935-942.

[12] YILDIRIM S, JOTHIMANI D, KAVAKLIOĞLU C, et al. Classification of "hot news" for financial forecast using NLP techniques[C]//Proceedings of the 2018 IEEE International Conference on Big Data. [S.l.]: IEEE, 2018: 4719-4722 .

[13] HEATON J B, POLSON N G, WITTE J H. Deep learning for finance: Deep portfolios[J]. Applied Stochastic Models in Business and Industry, 2017, 33(1): 3-12.

[14] 姚加权, 张锟澎, 罗平. 金融学文本大数据挖掘方法与研究进展[J]. 经济学动态, 2020(4): 143-158.
YAO J Q, ZHANG K P, LUO P. Text mining in financial big data and its research progress[J]. Economic Perspectives, 2020(4): 143-158.

[15] 岑咏华, 谭志浩, 吴承尧. 财经媒介信息对股票市场的影响研究: 基于情感分析的实证[J]. 数据分析与知识发现, 2019, 3(9): 98-114.
CEN Y H, TAN Z H, WU Z Y. Impacts of financial media information on stock market: An empirical study of sentiment analysis[J]. Data Analysis and Knowledge Discovery, 2019(9): 98-114.

[16] AKITA R, YOSHIHARA A, MATSUBARA T, et al. Deep learning for stock prediction using numerical and textual information[C]//Proceedings of the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). [S.l.]: IEEE, 2016: 1-6.

[17] EAPEN J, BEIN D, VERMA A. Novel deep learning model with CNN and bi-directional LSTM for improved stock market index prediction[C]//Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). [S.l.]: IEEE, 2019: 264-270.

[18] GAO T, LI X, CHAI Y, et al. Deep learning with stock indicators and two-dimensional principal component analysis for closing price prediction system[C]//Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). [S.l.]: IEEE, 2016: 166-169.

[19] SIM H S, KIM H I, AHN J J. Is deep learning for image recognition applicable to stock market prediction?[J]. Complexity, 2019(3): 1-10.

[20] VARGAS M R, DE LIMA B S, EVSUKOFF A G. Deep learning for stock market prediction from financial news articles[C]//Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). [S.l.]: IEEE, 2017: 60-65.

[21] 邹宗森, 杨素婷. 货币供应量、利率对汇率的影响——基于 VAR 模型的分析[J]. 金融教育研究, 2020, 33(3): 16-24.
ZOU Z S, YANG S T. The impact of money supply and interest rate on exchange rate: Analysis based on var model[J]. Research of Finance and Education, 2020, 33(3): 16-24.

[22] 王书芹, 华钢, 徐永刚, 等. AUC 的不一致性分析[J]. 江苏师范大学学报 (自然科学版), 2013, 31(3): 31-34.
WANG S Q, HUA G, XU Y G, et al. The incoherence of the area under the ROC curve[J]. Journal of Jiangsu Normal University (Natural Science Edition), 2013, 31(3): 31-34.

编 辑 叶 芳