Finance 金融, 2021, 11(6), 535-546
Published Online November 2021 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2021.116058

Hans汉斯

基于财经新闻V-A情感分歧的股票价格预测方法研究

宋　莹，马　彪

东华大学旭日工商管理学院，上海

收稿日期：2021年10月12日；录用日期：2021年11月5日；发布日期：2021年11月12日

摘　要

互联网财经新闻已然成为股市投资者获取股票相关信息的首要来源，其引发的投资者情感波动必然会对股市造成影响，充分挖掘财经新闻潜在的情感信息，可以更好地洞察股票市场趋势。而现有的财经新闻情感分析，常忽略新闻情感信息的多样性，只考虑单维度情感，无法量化复杂情感隐含状态，造成情感信息缺失。因此，本文旨在利用财经新闻板块V-A多维度情感分歧对股票市场进行研究，提高股票板块价格预测精度。利用CNN-LSTM组合模型提取文本局部特征与语义特征，构建财经新闻连续维度V-A情感计算模型，从连续多维度量化复杂情感，更全面地表示情感信息，进而精确地计算财经新闻情感分歧。后结合情感分歧与股票价格构建股票板块价格预测模型，采用GridSearchCV对SVR预测模型进行参数寻优。模型预测准确率进一步提升，其中，制造业板块平均绝对误差为0.2030，证明本文V-A连续计算情感模型测度新闻情感，可以有效地提高股票预测准确率。

关键词

连续维度V-A情感模型，情感分歧，支持向量回归

Stock Price Forecasting Based on V-A Sentiment Divergence of Financial News

Ying Song, Biao Ma

Glorious Sun School of Business and Management, Donghua University, Shanghai

Received: Oct. 12th, 2021; accepted: Nov. 5th, 2021; published: Nov. 12th, 2021

Abstract

Internet financial news has become the primary source for stock market investors to obtain stock-

文章引用: 宋莹, 马彪. 基于财经新闻 V-A 情感分歧的股票价格预测方法研究[J]. 金融, 2021, 11(6): 535-546.
DOI: 10.12677/fin.2021.116058

related information. Investors' sentiment fluctuations triggered by financial news will inevitably affect the stock market. Fully mining the potential sentiment information of financial news can provide better insights into stock market trends. However, prior financial news sentiment analysis methods often consider single-dimensional sentiment, ignore the diversity of news sentiment information, and cannot quantify the complex sentiment, resulting in a lack of sentiment information. Therefore, this article aims to use the V-A multi-dimensional sentiment divergence of the financial news sector to study the stock market and improve the accuracy of stock sector price forecasts. The CNN-LSTM combined model is used to extract the local and semantic features of the text, and the continuous-dimensional V-A sentiment computing model of financial news is constructed, which quantifies complex sentiment from continuous multi-dimensional and expresses sentiment information more comprehensively, so as to accurately compute the sentiment divergence of financial news. After combining sentiment divergence and stock prices, the stock sector price prediction model is constructed, and GridSearchCV is used to optimize the parameters of the SVR prediction model. The accuracy of model prediction is further improved. Among them, the average absolute error of the manufacturing sector is 0.2030, which proves that the V-A continuous computing sentiment model to measure news sentiment can effectively improve the accuracy of stock prediction.

Keywords

Continuous Dimension V-A Emotional Model, Sentiment Divergence, SVR

1. 引言

互联网财经新闻，作为一种公开发布并且具有极强公信力的媒体信息，提供了股票市场的国家政策、行业信息等各方面内容[1]，已然成为股市投资者获取股票相关信息的首要来源。财经新闻引发的投资者情感波动必然会对股市造成影响。充分挖掘财经新闻潜在的情感信息，可以更好地洞察股票市场趋势。

因此，越来越多的学者关注金融文本情感与股票波动的关系，但目前基于金融文本情感的研究多集中于个股。而行业因素历来也是投资者考察股市的重要因素之一，行业板块的波动很大程度表现出行业的风险[2]。因此，股票市场板块波动程度与新闻情感的关系更值得关注。另一方面，目前基于新闻情感的股市研究，常将平均情感作为情感参数[3] [4]。然而平均情感隐藏了很大的情感变化，忽略了新闻情感信息的多样性，造成大量信息损失，而情感分歧可以准确地描述情感的多样性[5]。因而有必要准确计算情感分歧，将情感分歧引入股票预测模型，从而能够更充分地挖掘财经新闻情感信息，减少信息损失。

同时，现有新闻情感分析未考虑情感的粒度划分，只考虑单个维度的情感信息，无法量化复杂情感的隐含状态，导致情感分歧计算的情感信息缺失。而连续维度情感模型，将不同情感根据维度的属性分布在空间中不同位置，实现情感细粒度划分，以多维度连续数值量化复杂情感，区分交叉和重叠的情感，更全面地表示情感信息，进而更精确地计算财经新闻情感分歧。

本文旨在研究财经新闻板块情感分歧对股票板块价格影响，从深入挖掘财经新闻情感角度出发，构建多维情感模型将财经新闻转化为影响股票市场波动的情感指标，确定新闻情感分歧计算方法，结合板块情感分歧与股票数据构建股票板块预测模型，提升股票板块预测准确率，为股市投资者提供可靠的决策支持。

2. 文献综述

随着股票市场的发展和财经新闻的不断普及，越来越多的学者考虑财经新闻对股市价格的影响。徐伟[6]利用情感分析量化财经新闻，研究与股票的关系，发现股票预测准确率受到新闻数量和质量的影响。赵澄等人[7]提出针对海量新闻的多维情感特征向量化方法，利用支持向量机(SVM)预测金融新闻对股票市场的影响。关于股票回归预测，冉杨帆[8]结合情感分析和机器学习，分别采用 BPNN 和 SVR 进行股票预测。Jheng-Long 等人[3]利用新闻情感预测股票走势，发现将财经新闻情感指标和股票指标都纳入预测特征，可以提高预测的准确度。Bollen 等[9]从 6 个维度对大量的 Twitter 消息进行情感分析，通过自回归模糊神经网络预测道琼斯指数的收盘价。然而，随着研究的不断深入，发现了部分互相矛盾的结果。如 Robert 等人[4]发现正面文本容易预测股票价格下降，负面和中性文本更容易预测价格上升，而这一结果与学者 P. Tetlock [10]等人观察到负面情感表明价格下跌的结论产生了矛盾。矛盾的产生不仅与语料库相关，更与财经新闻情感分析方法有关，研究中纳入的情感信息多样性会对预测结果造成影响。

在情感强度指标上，目前的研究多选择平均情感强度。然而，平均情感强度会忽略情感多样性，造成情感信息损失，如情感中立的平均情感与一半积极一半消极的结果一致，进而影响预测模型精确度。Siganos 等人也提到平均情感掩盖了情感的重要差异，情感分歧可以更充分地描述情感信息，与股价波动呈正相关，并表明较高的分歧会导致较高的绝对股票价格变动[5]。但是 Siganos 等人的情感分歧方法仅仅通过积极与消极词汇的词频统计来描述，未考虑文本局部特征及语义特征，无法很好地描述情感信息。也有学者[11] [12]利用情感词典确定情感词性或情感程度级别，考虑情感词修饰结构等计算情感分歧。这些方法的准确度深受情感词典完善程度影响，而且不能区分不同词汇情感强度，无法量化复杂情感，也会对预测模型效果造成影响。

情感状态的表示上，主要有两种类型[13]：离散型和连续维度型。现有研究多为离散型，将情感划分为若干离散情感标签，并运用文字、语言等特征分类[14]。连续维度型，指情感分布在若干维度组成的空间，不同情感根据维度的属性分布在空间中不同位置，以连续数值量化情感[15]。相较于离散型，连续型可以从多维量化复杂情感，区分交叉和重叠的情感，更全面的表示情感信息，进而更精确的计算财经新闻情感分歧。目前较为常见的连续维度研究为 Valence-Arousal 二维空间模型[16]，将文本情感映射到二维情感空间，V 为愉悦度 Valence，表示情感的正面负面程度，A 是激活度 Arousal，表示情感的平静与激动程度，两个维度均为[0, 9]的连续实数值。目前 V-A 二维情感分析多为基于情感词典计算相似度或者直接将词语特征输入神经网络组合模型进行计算，Jin Wang [17]基于区域划分文本，利用 CNN-LSTM 模型考虑句子的本地信息和句子之间的长距离依赖，分析中英文文本的连续维度型情感。Chuhan Wu [18]等人提出了一种基于变分自编码器模型的半监督 LSTM 模型进行句子文本 V-A 二维情感计算。胡佳男[19]提出基于线性加权的 W-CNN-LSTM 模型进行句子文本 V-A 二维连续情感计算。

综上所述，本文旨在研究财经新闻板块情感对股票板块市场的影响，引入 V-A 连续二维情感模型考虑新闻文本局部特征及语义特征，分析财经新闻情感，并纳入包含丰富信息的情感分歧构建 SVR 股票板块预测模型，以期提高股票板块价格预测精度。

3. 模型构建

本文考虑财经新闻文本局部情感信息和语义特征，构建更精确的 V-A 二维情感计算方法，分析新闻情感确定包含丰富信息的新闻情感分歧，将财经新闻转化为影响股票市场趋势的情感指标，并结合股票信息构建 SVR 股票板块预测模型，以提高股票预测准确率。财经新闻板块情感分歧的股票预测研究主要包括三部分：1) 财经新闻板块标签，将新闻按板块划分，为股票板块预测做准备。2) 新闻板块情感分歧，

结合 V-A 连续情感模型，考虑词汇局部特征及语义特征，挖掘新闻情感信息，确定情感分歧。3) 股票预测模型，将情感分歧和股票指标作为输入变量，并对 SVR 模型进行参数寻优，构建预测模型。如图 1。

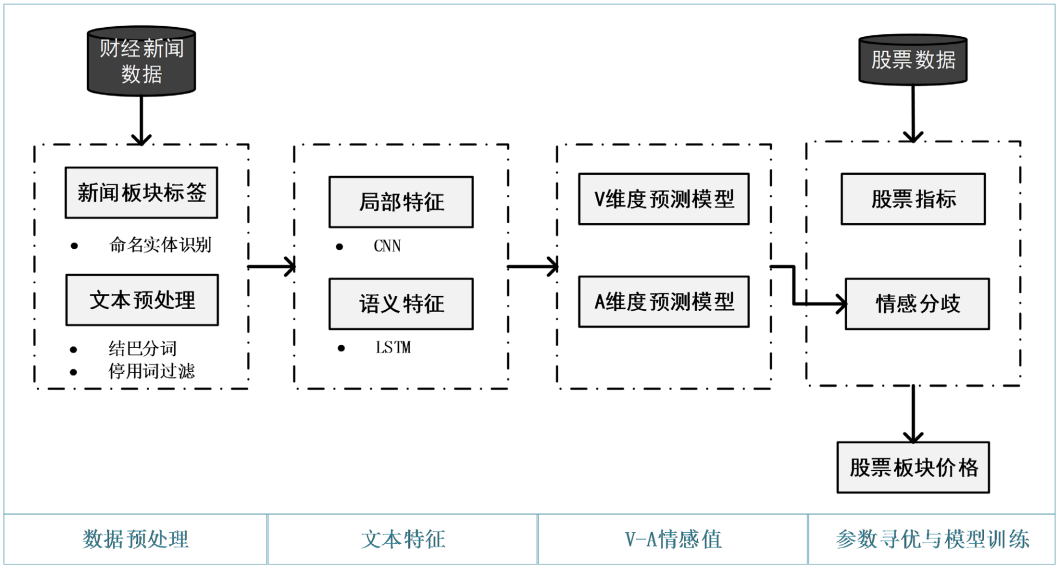


Figure 1. Framework for stock forecasting based on sentiment divergence of financial news
图 1. 财经新闻板块情感分歧的股票预测框架

3.1. 财经新闻 V-A 情感强度计算模型

财经新闻的 V-A 二维连续情感强度由新闻语句的文本局部特征和语义特征决定。因此，仅仅将文本情感词数量等信息作为情感特征，在表达发生变化时，如变为否定句、疑问句，其准确率很难保证，需要考虑词汇的局部特征。深度神经网络多用于文本的深层建模与特征学习，能够获取大量数据中隐藏信息，构建复杂模型。卷积神经网络(Convolutional Neural Networks, CNN)利用卷积核提取词汇与其上下文的关系，无需人为干预自动学习文本的隐藏特征。因此，考虑利用 CNN 提取财经新闻局部特征信息。进而考虑到财经新闻的文字表达形式自由但结构上存在上下文依赖关系。其语义特征一定程度上描述了句子文本的上下文信息，能够更准确地理解文本语义，符合人的主观意识。考虑到 CNN 卷积核的特殊性，无法捕获到文本的长距离依赖关系，只能捕获到局部特征信息，本文引入长短期记忆网络(Long Short-Term Memory, LSTM)获取长距离的语义信息，学习句子之间的依赖关系，减少情感语义的丢失，捕获财经新闻文本语义特征。

综上所述，文本局部特征和语义特征会影响财经新闻情感表达，为了同时提取文本局部信息和语义信息，采用 CNN-LSTM 组合模型分别计算新闻文本 VA 两个维度的情感值，得到新闻情感强度。由于文本级别句子关系复杂、语义信息多样，很难准确地进行情感分析，而文本通常由多个句子构成，将文本情感转化为句子层级的情感分析，可以准确的描述情感。因此，将新闻句子文本向量作为模型输入，利用 CNN 模型提取词汇局部特征，LSTM 用于提取句子逻辑关系，最后利用线性激活函数整合计算句子文本 V-A 情感值。财经新闻 VA 二维情感强度测度框架如图 2 所示，包括嵌入层、CNN Layer 部分的卷积层和最大池化层、LSTM Layer 部分以及线性激活函数。

(1) 词嵌入部分

新闻句子文本的长度是变化的，无法应用于预测模型，同时为了捕获语义和句法信息，需要针对句

子文本采用词嵌入技术，从更高的语义层面描述词与词之间的关系。词嵌入部分将句子sentence从单词序列$[w_1, w_2, \cdots, w_n]$转换为低维密集的向量序列$[e_1^w, e_2^w, \cdots, e_n^w]$。

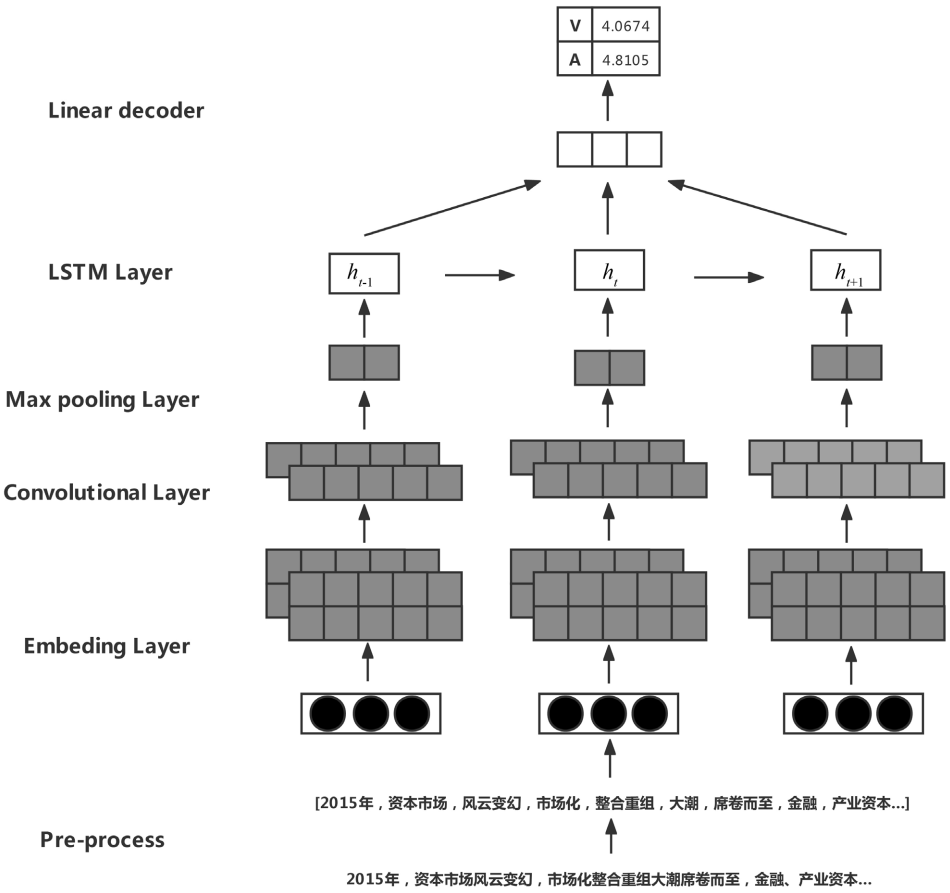


Figure 2. Frame of VA sentiment intensity measurement based on CNN-LSTM model
图 2. 基于 CNN-LSTM 模型的 VA 情感强度测度框架

(2) CNN Layer 部分

财经新闻句子文本向量作为 CNN 层输入，CNN Layer 作为神经网络模型的隐含层部分，包括卷积层(Convolutional Layer)和池化层(Pooling Layer)。

1) 卷积层

卷积神经网络可以方便的利用卷积核尺寸，提取句子中每个词与其上文和下文中的关系。通过卷积核进行卷积类似于语言建模，可以有效的提取特征，减少无关信息的影响。句子文本情感卷积操作中，利用卷积核提取局部 n-gram 情感特征。

2) 最大池化层

池化层整合卷积产生的特征映射，降低中间隐含层的维度，避免过拟合，同时增强局部统计特征的感受，提高平移不变性。本文采用最大池化层，去除非最大值的特征降低上层计算量，并且可以提取句子局部依赖关系保持局部情感信息。池化层的向量最终被输入 LSTM Layer。

(3) LSTM Layer 部分

长短时记忆网络模型通过遗忘门、输入门、输出门保存区域之间的逻辑信息，很好地捕获句子的长

依赖关系。本文将 CNN Layer 部分卷积核获取的特征映射经过池化层采样后提取特征，作为 LSTM Layer 的输入，通过 LSTM 的隐含层结构，进一步提取卷积层输出的特征之前的顺序关系。

(4) 线性激活函数

神经网络方法多用于分类任务，而 VA 情感强度测度旨在获取的两个维度情感为连续实数值，CNN-LSTM 无法直接应用于本文的 V-A 连续维度情感强度计算，因此采用线性激活函数进行转换，获得单一实数值输出，作为财经新闻句子情感强度。该线性激活函数可以表示为：

$$y = Wx + b \tag{1}$$

其中，x 为 LSTM Layer 部分学习获得的句子向量，y 表示目标文本的 Valence 或 Arousal 情感值，W 表示线性激活函数的权重，b 表示线性激活函数的偏置量。

最终将 V、A 两个维度的计算的结果整合，得到新闻句子的 VA 二维情感强度：

$$e = \sqrt{v^2 + a^2} \tag{2}$$

则该篇财经新闻情感强度为 $s = \frac{\sum_{d=1}^{D} e_d}{D}$，其中 D 为该篇财经新闻所拆分的文本句子数量。

3.2. 财经新闻板块情感分歧计算

(1) 情感分歧

基于财经新闻情感的股票预测模型效果很大程度上取决于新闻情感的表示方式，因为股票的价格波动会受到多种情感因素影响，纳入情感因素的丰富程度是预测模型准确度高低的关键。因此，考虑新闻的股票预测模型要求纳入更全面的情感信息，以得到更精确的结果。而情感分析常用的平均情感很大程度上忽略了情感的多样性与差异性，造成情感信息损失，无法充分地刻画情感。信息论角度来看，香农 1948 年提出[20]信息熵的概念，用来描述信息的不确定程度，常被引用计算分歧度。基本原理是，情感值不确定度越高，其情感差异越大。因此，若某天出现大量情感各异的财经新闻，其不确定性较高，包含较高的信息量。衡量情感差异的情感分歧可以更精确地描述情感的不确定，减少情感信息损失。本文的“情感分歧”指在财经新闻不同报道源对某一天事件或话题所表达出的情感差异性。情感各异新闻数据较多，也会产生较大的情感分歧，得到较高的情感信息熵。

Siganos 等人[5]利用第 i 天评论正面与负面词的词频统计均值与总均值的差距衡量该天情感分歧。该方法适用于观察特定时间情感的差异，但仅仅基于词频统计，未考虑文本特征，忽略了文本表达的情感强烈程度，无法充分刻画情感。因此，为避免情感信息损失，本文基于 3.1 中建立的连续维度 V-A 情感计算模型，重新构建情感分歧，充分描述情感的不确定。

(2) 板块情感分歧计算方法

V-A 情感计算模型得到财经新闻句子文本的情感强度 s_j，情感强度计算得到的为[0, 9]的正向连续情感值，为区分积极与消极情感，将新闻情感强度归一化得到[−1, 1]的连续实数值，财经新闻情感强度为：

$$senti_j = \frac{s_j - \bar{s}}{\sigma_s} \tag{3}$$

其中，$\bar{s}$、σ_s 分别为该股票板块该段时间的情感均值与方差。将情感强度 $senti_j$ 为正值的判定为积极情感，负值的为消极情感。则第 i 天某板块积极财经新闻情感强度为 $x_{p,i} = \sum_{j=1}^{P} \frac{senti_{p,j}}{Z}$，$P$ 表示第 i 天股票板块的积极新闻篇数，同理，第 i 天某板块消极新闻情感强度为 $x_{n,i} = \sum_{j=1}^{N} \frac{senti_{n,j}}{Z}$，$N$ 表示第 i 天板块的消极

新闻篇数。

则第 i 天某股票板块积极情感指标为

$$\left|\frac{x_{p,i}-x_{p,all}}{\sigma_{p,all}}\right| \tag{4}$$

同理，消极情感指标为

$$\left|\frac{x_{n,i}-x_{n,all}}{\sigma_{n,all}}\right| \tag{5}$$

若某天正负情感指数都很高，表明这天存在大量积极与消极信息，则情感分歧度也将较高。如果大量积极和较少消极状态，情感分歧将相对较低。因此，情感分歧为

$$DOS_i=\left|\frac{x_{p,i}-x_{p,all}}{\sigma_{p,all}}\right|+\left|\frac{x_{n,i}-x_{n,all}}{\sigma_{n,all}}\right| \tag{6}$$

其中，$x_{p,i}$，$x_{n,i}$：第 i 天某板块积极与消极新闻情感强度均值，$x_{p,all}$，$x_{n,all}$：时间段内某板块情感强度均值，$\sigma_{p,all}$，$\sigma_{n,all}$：时间段内某板块情感强度方差。DOS_i 的值越大则表示该天该板块的情感分歧越大。

3.3. 财经新闻情感分歧的股票价格预测模型

较高的情感分歧可能导致较高的绝对价格变化[21]，Siganos 发现情感分歧与股价波动呈正相关，但未对股票价格如何随情感分歧变化进行研究。本文以数值形式揭示新闻情感分歧对股票板块价格的影响，因此将新闻板块情感分歧作为股票预测模型的情感参数。许多学者发现 SVR 模型预测股票价格误差更小，同时，SVR 可以解决传统神经网络采用经验风险最小化而导致的过拟合的问题，在非线性、小样本的训练中具有独特的优势[22]。考虑到本文的预测模型输入的非线性等特征，选择利用 SVR 构建股票预测模型，更好的挖掘新闻情感分歧与股票价格的关系。

按照约定俗成把个股收盘价作为当日个股股票价格。同时，Jheng-Long 等人[3]发现将财经新闻情感指标和股票指标都纳入预测特征，可以提高预测准确度。因此，将前一期财经新闻板块情感分歧 DOS_i 与股票板块价格 $Stock_i$ 作为特征纳入预测模型训练。训练样本的特征为 $x_i=DOS_i,Stock_i$，输出变量为 $y_i=\langle Price_i\rangle$。SVR 预测模型通过非线性映射 $x\to\varphi(x)$，将非线性关系样本组x投影到高维特征空间成为线性关系，后在高维特征空间进行线性回归，高维特征空间的线性模型为：

$$f(x)=w\cdot\varphi(x)+b \tag{7}$$

$\varphi(x)$ 为非线性映射，w 为权向量，b 为阈值。SVR 寻求的最优超平面是所有样本点离超平面总偏差最小，则可以描述为优化问题：

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\ell_\varepsilon\left(f(x_i)-y_i\right)\\&s.t.y_i\left(w^T\varphi(x_i)+b\right)\geq 1,i=1,2,\cdots,m\end{aligned} \tag{8}$$

C 为惩罚因子，ℓ_ε 是 ε 不敏感损失函数。同时，考虑到股票数据的非线性和复杂性，选取高斯核函数完成特征空间映射，核函数表示为：

$$K(x,y)=\exp\left(-\gamma\|x-y\|^2\right) \tag{9}$$

C 和 *gamma* 是影响 SVR 模型泛化性能的重要参数，其设定对训练效果有最显著的影响。参数 *C* 是模型训练的惩罚系数，即模型训练误差的宽容度。*gamma* 决定数据映射到高维特征空间后的分布，调整模型的复杂度。由于待调节参数之间的组合繁复，本文利用参数自动搜索(GridSearchCV)，在指定的范围内自动搜索不同参数的模型组合，获得最优的参数以使得 SVR 模型性能最优。最终，本文 SVR 模型的回归函数可以表示为：

$$f(x)=\sum_{i=1}^{m}\left(\alpha_i-\alpha_i^*\right)\exp\left(-\gamma\left\|x_i-x\right\|^2\right)+b \tag{10}$$

其中，α_i 和 α_i^* 分别为拉格朗日函数和乘子。

4. 实验分析

4.1. 数据来源及处理

(1) 数据来源

选取搜狐财经 2016 年 1 月到 2016 年 4 月的财经新闻数据作为研究语料库，同时删除部分重复出现、不规范的数据，共计 4820 条财经新闻数据。

V-A 二维情感计算的情感语料库选取元智大学自然语言处理实验室的中文维度型情感语料库(Chinese Valence-Arousal Words, CVAW)，其中新闻句子文本共计 1509 条。

选用的股票交易数据来源于锐思金融研究数据库，该数据库是国内目前数据涉及最全面、用户使用最方便的研究数据库。

(2) 数据预处理

财经新闻为中文自然语言，是非结构化的复杂文本形式，需要对新闻文本进行预处理，得到有意义的词序列。首先采用 jieba 分词的精确模式将一系列的中文句子切分成词语序列。其次过滤停用词，本文建立专用停用词表完成停用词过滤，包括无意义词汇、特殊符号等，共计 1928 个词汇。

(3) 财经新闻的板块标签

行业因素历来是投资者考察股市的重要因素之一，行业板块的波动很大程度表现出行业的风险[2]，股票市场板块波动程度与新闻情感的关系更值得关注。因此，本文以新闻板块情感分歧对股票板块波动进行预测。选用哈工大的 LTP 命名实体识别技术识别公司名称，确定财经新闻板块标签。为使公司名称分词结果精确，加载自定义的包含简称及全称的公司名称词典进行分词。LTP 提供的命名实体类型为：人名(Nh)、地名(Ns)、机构名(Ni)。针对“Ni”词性的词汇进一步的识别，根据词汇位置，是实体中间词、实体结束词还是单独成实体等进一步划分，从而识别新闻公司名称并确定其板块标签。按证监会行业分类将财经新闻与股票数据划分为 19 个板块。

4.2. 财经新闻 V-A 情感计算模型结果

(1) V-A 情感模型计算结果

应用 CNN-LSTM 模型充分挖掘新闻文本局部特征及语义信息，分别计算 Valence、Arousal 连续二维情感值，并将 V、A 两个维度的测度的结果整合，得到新闻文本情感强度，部分结果如表 1 所示。

(2) 新闻情感分歧计算结果

上文获得财经新闻情感强度，通过归一化得到每天的新闻板块的积极与消极情感，取值范围为[−1, 1]。最终，根据积极、消极情感计算得到每天新闻板块的情感分歧指标，部分结果如表 2 所示。

Table 1. V-A Sentiment intensity computing results of financial news
表 1. 财经新闻 V-A 情感强度计算结果

NewsCode	News	Valence	Arousal	Senti
1	本周一周二，沪指企稳反弹迹象明显，这为题材股表演提供了舞台…	4.4163	4.3249	6.1894
2	在披露重大资产重组的进展不久，万科再次发出公告提示其重组事项存在的不确定性…	4.4094	4.2809	6.1582
3	悄悄告诉你个新动向，新股回拨至少可 67%至 200%提升网上中签率…	4.3905	4.3333	6.1799
4	根据中国证券监督管理委员会《关于进一步规范证券投资…	3.9712	4.3633	5.9120
5	2015 年，资本市场风云变幻，市场化整合重组大潮席卷而至，金融、产业资…	4.4683	4.3448	6.2435

Table 2. Sentiment divergence of financial news
表 2. 每天财经新闻板块情感分歧

PlateCode	PlateName	Date	Senti_Div
C	制造业	2016/1/4	1.2310
C	制造业	2016/1/5	0.1074
C	制造业	2016/1/6	1.7674
C	制造业	2016/1/7	1.3598
C	制造业	2016/1/8	2.2841

4.3. 财经新闻板块情感分歧的股票价格预测结果

股票板块预测模型构建，将前一期财经新闻板块情感分歧与前一期股票板块价格作为模型输入，选用 GridSearchCV 确定最优 SVR 模型的参数 C 与 $gamma$，以输出该期股票板块预测价格。选择平均绝对误差(MAE)和均方误差(MSE)衡量预测模型效果。由于财经新闻标题是全文表达的概述，亦可以表现新闻情感，考虑分别将新闻全文和标题情感分歧作为股票预测的情感参数，观测效果。

实验以 C 板块即制造业为例预测股票板块价格，选取 2016 年 1 月到 3 月的财经新闻及股票数据为训练集，2016 年 4 月的数据为测试集。SVR 预测模型通过网格搜索进行参数寻优，得到最优参数组合 $\{C:1.0, gamma:0.1\}$。另外，为验证本文财经新闻情感模型的有效性，对比其他方法进行预测。如表 3。

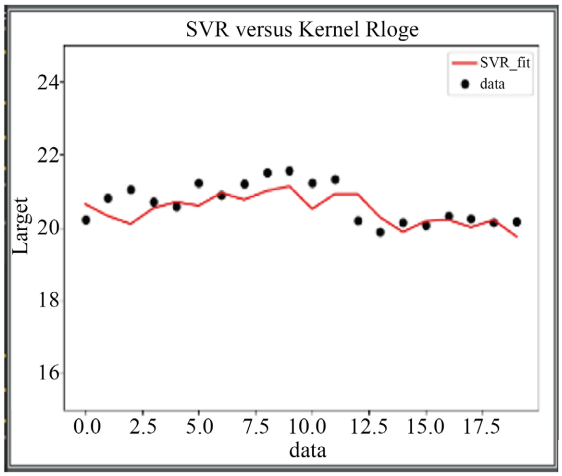
Table 3. Comparison of the Prediction Results of Plate C
表 3. 板块 C 股票价格预测结果对比

股票预测模型输入		MSE	MAE
情感分歧 + 股票指标融合$_{[EDS]}$	Train	0.5085	0.5161
	Test	0.2030	0.3829
情感倾向 + 股票指标融合$_{[ES]}$	Train	0.4272	0.4906
	Test	0.3344	0.3849
新闻标题情感分歧 + 股票指标$_{[TEDS]}$	Train	0.6521	0.5573
	Test	0.2425	0.3832

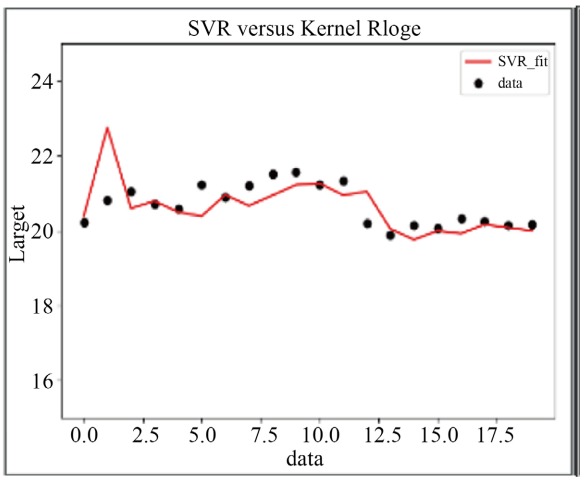
Continued

情感分歧	Train	1.7220	1.0417
	Test	2.9512	1.6322
股票指标	Train	0.6649	0.5651
	Test	0.2515	0.3703

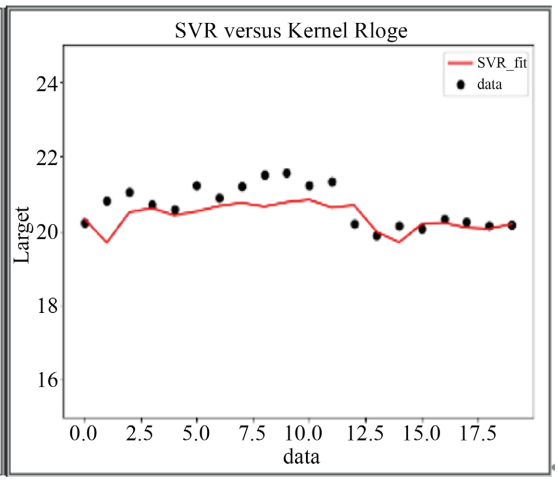
[EDS]为利用本文 V-A 情感模型计算新闻全文情感分歧，构建的股票预测模型，测试集的 MSE、MAE 为 0.2030、0.3829，对比其他预测方法准确度相对较高；[ES]为 Jheng-Long [3]于 2012 年提出的经典的基于情感强度的预测模型，本文将 V-A 情感值作为情感参数加以实现；[TEDS]考虑新闻标题情感是否可以更好地表示情感，对比新闻全文较差；除此之外对比单一的预测模型输入，即只包括股票指标或情感分歧，得到结果 MSE、MAE 都比本文提出的预测方法大。从对比结果来看，本文的方法可以有效提高股票板块价格预测精度。[EDS]、[ES]和[TEDS]测试集的预测结果图 3 所示。



情感分歧+股票指标



情感倾向+股票指标



标题情感分歧+股票指标

Figure 3. Plate C testing set stock price prediction results
图 3. 板块 C 测试集股票价格预测结果

5. 结语

随着网络上涌现越来越多的财经领域新闻，互联网财经新闻能造成股票市场的波动已经是公认的事实，如何从浩瀚如海的财经新闻中获取潜在情感信息，以洞察股票市场趋势愈来愈成为规避股票市场投资风险的关键。基于此，本文提出基于互联网财经新闻情感分歧的股票板块预测模型，聚焦于股票板块，考虑新闻文本局部特征及语义特征，计算财经新闻的 V-A 二维连续情感强度，并确定财经新闻情感分歧，再结合板块情感分歧与股票指标运用 SVR 完成股票板块预测模型构建，提升了股票板块预测准确率，为股市投资者提供准确可靠的决策支持以规避投资风险。

本文的研究涉及了文本挖掘，机器学习，金融分析等多个学科领域，预测模型取得了一定成果，但还有不足之处，需要在未来研究中进一步改进：(1) 目前选用有监督多通道的 CNN-LSTM 方法计算财经新闻情感强度，依赖语料库影响情感分析精准度，接下来可利用半监督方式进一步优化。(2) 文本情感的表达受多种文本特征影响，而本文目前只考虑了词语语义特征，可以纳入更多文本特征以更充分地描述文本情感。

基金项目

本文系中央高校基本科研业务费专项资金资助项目(项目编号：2232018H-07)的研究成果之一。

参考文献

[1] 刘欣. 互联网财经新闻媒体对中国股市的影响力排名研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2014.

[2] 柳燕燕. 中国股票市场行业板块波动研究[D]: [硕士学位论文]. 抚州: 东华理工大学, 2013.

[3] Wu, J.-L., Su, C.-C., Yu, L.-C., *et al*. (2012) Stock Price Predication Using Combinational Features from Sentimental Analysis of Stock News and Technical Analysis of Trading Information. In: *Proceedings of International Conference on Economics, Business and Management* (*ICEBM*2012), IEDRC: Chengdu Young Education & Consultancy Co., Chengdu, 5.

[4] Schumaker, R.P., Zhang, Y.L., Huang, C.-N., *et al*. (2012) Evaluating Sentiment in Financial News Articles. *Decision Support Systems*, **53**, 458-464. https://doi.org/10.1016/j.dss.2012.03.001

[5] Siganos, A., Vagenas-Nanos, E. and Verwijmeren, P. (2017) Divergence of Sentiment and Stock Market Trading. *Journal of Banking & Finance*, **78**, 130-141. https://doi.org/10.1016/j.jbankfin.2017.02.005

[6] 徐伟, 李韵喆. 行业与个股新闻对股票价格影响的定量分析[J]. 财经界(学术版), 2015(13): 31-32.

[7] 赵澄, 叶耀威, 姚明海. 基于金融文本情感的股票波动预测[J]. 计算机科学, 2020, 47(5): 79-83.

[8] 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.

[9] Bollen, J., Mao, H.N. and Zeng, X.J. (2011) Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, **2**, 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

[10] Tetlock, P.C. (2007) Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, **62**, 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

[11] 吕华揆, 刘政昊, 钱宇星, 洪旭东. 异质性财经新闻与股市关系研究[J/OL]. 数据分析与知识发现, 1-18. http://kns.cnki.net/kcms/detail/10.1478.G2.20200907.0945.002.html, 2020-12-20.

[12] 徐健, 吴思洋. 网络用户评论的情感分歧度量化算法研究[J]. 情报学报, 2020, 39(4): 427-435.

[13] Calvo, R.A. and Mac Kim, S. (2013) Emotions in Text: Dimensional and Categorical Models. *Computational Intelligence*, **29**, 527-543. https://doi.org/10.1111/j.1467-8640.2012.00456.x

[14] 王津. 基于 Valence-Arousal 空间的中文文本情感分析方法研究[D]: [博士学位论文]. 昆明: 云南大学, 2016.

[15] Bakker, I., Voordt, T., Vink, P. and Boon, J. (2014) Pleasure, Arousal, Dominance: Mehrabian and Russell Revisited. *Current Psychology*, **33**, 405-421. https://doi.org/10.1007/s12144-014-9219-4

[16] Russell, J.A. (1980) A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, **39**, 1161-1178. https://doi.org/10.1037/h0077714

[17] Wang, J., Yu, L., Lai, K.R. and Zhang, X. (2020) Tree-Structured Regional CNN-LSTM Model for Dimensional Sen-

timent Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 581-591. https://doi.org/10.1109/TASLP.2019.2959251

[18] Wu, C.H., Wu, F.Z., Wu, S.X., Yuan, Z.G., Liu, J.X. and Huang, Y.F. (2018) Semi-Supervised Dimensional Sentiment Analysis with Variational Autoencoder. *Knowledge-Based Systems*, **165**, 30-39. https://doi.org/10.1016/j.knosys.2018.11.018

[19] 胡佳男. 基于连续维度型的文本情感强度计算方法研究[D]: [硕士学位论文]. 南昌: 南昌大学, 2017.

[20] Shannon, C.E. (2001) A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**, 3-55. https://doi.org/10.1145/584091.584093

[21] 孙秋韵, 刘金清, 刘引, 等. 基于改进 GA 参数优化的 SVR 股价预测模型[J]. 计算机系统应用, 2015, 24(9): 29-34.

[22] 周涛丽. 基于支持向量机的多分类方法研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.