Computer Science and Application 计算机科学与应用, 2024, 14(1), 88-97
Published Online January 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.141010

Hans汉斯

面向图片和文本多源异质数据的股票预测联合模型

昝泓含，李艳艳

燕山大学，理学院，河北 秦皇岛

收稿日期：2023年12月22日；录用日期：2024年1月19日；发布日期：2024年1月30日

摘　要

股票预测一直是金融界研究的热点问题，近年来融合文本、图片这类非结构化数据成为提高预测精度的研究方向。本文建立了一种能够同时处理多源异质数据的股票价格走势预测联合模型，分析词云图片、股吧评论文本和股票交易数据。联合模型分为两个分支，一个分支运用CNN模型分析由股民评论文本转为的词云图片，另一分支运用LSTM模型处理历史股票交易数据和由股民评论文本得到的情感评分，两个分支共同预测4天、6天、8天的股票走势涨跌。结果表明使用词云图片的CNN模型表现优于情感分析的LSTM模型，证明词云图片的可使用性，且联合模型结果优于两个单一模型，准确率稳定在0.6~0.7之间。

关键词

多源异质数据，词云图片，情感分析，LSTM，CNN

A Joint Model for Stock Prediction Based on Image and Text Multisource Heterogeneous Data

Honghan Zan, Yanyan Li

School of Science, Yanshan University, Qinhuangdao Hebei

Received: Dec. 22nd, 2023; accepted: Jan. 19th, 2024; published: Jan. 30th, 2024

Abstract

Stock prediction has always been a hot topic in the financial industry, and in recent years, inte-

文章引用: 昝泓含, 李艳艳. 面向图片和文本多源异质数据的股票预测联合模型[J]. 计算机科学与应用, 2024, 14(1): 88-97. DOI: 10.12677/csa.2024.141010

grating unstructured data such as text and images has become a research direction to improve prediction accuracy. This article establishes a joint model for predicting stock price trends that can simultaneously process multi-source heterogeneous data, including the analysis of word cloud images, stock bar comments, and historical stock trading data. The proposed model comprises two branches. The first branch utilizes a CNN model to dissect word cloud images derived from stock comment texts, while the second branch employs an LSTM model to process historical stock trading data and emotional scores gleaned from the stock comment texts. Two branches jointly predict the rise and fall of stock trends for 4, 6, and 8 days. The findings indicate that the CNN model's use of word cloud images yields superior performance compared to the LSTM model's sentiment analysis. This outcome underscores the efficacy of leveraging word cloud images as a predictive tool. Moreover, the joint model's results surpass those of the individual models, with an achieved accuracy consistently ranging between 0.6 and 0.7.

Keywords

Multi Source Heterogeneous Data, Word Cloud Images, Emotional Analysis, LSTM, CNN

1. 引言

股票预测一直是金融界学者们研究的热点问题。随着互联网的发展和技术的进步，学者们在输入数据时，开始融合结构化数据和非结构化数据，目前使用较多的非结构化数据为图片、文本两类。使用这些多源异质数据，能够优化数据输入的多元性，使模型获取的市场信息更加全面，从而提高预测精度[1]。

目前学者们对于文本类数据的研究分为两类，即金融新闻类文本和社交媒体类评论文本。Maqbool等(2023) [2]使用多层感知器回归模型(MLP-regressor)分别对 10 天、30 天和 100 天的股票进行预测，检验金融新闻情绪对股票走势的影响。结果表明股价与金融新闻之间存在高度相关性，使用金融新闻情绪可以预测股价。Liu 等(2020) [3]从股吧获取股票评论文本并与股票历史交易数据一起进行股票预测分析，结果表明评论文本数据和股票历史交易数据是相辅相成，两者可用作提取互补特征的数据源。

学者们对于文本类数据的使用逐步成熟，但对于图片数据的研究仍处于起步阶段。现在普遍使用的图片是蜡烛图，蜡烛图内包含股票价格、走势信息，能够取代股票历史交易数据这类结构化数据。Hung 和 Chen (2021) [4]通过蜡烛图预测 NI225 指数 2015~2019 年价格涨跌趋势，平均准确率达到 66.53%。学者们认为对于股票这种容易受到社会新闻事件、相关行业股票、投资者情绪等多种因素影响的敏感型投资产品，利用知识图谱是一种很有前景的研究方向。知识图谱能将与股票相关联的信息融入到价格特征表示中，从而增加多因素输入，使得模型能学习到更多市场信息，进而提高预测精度。可是，目前知识图谱研究工作具有大量节点的图的复杂性和梯度问题有待解决。Jafari 和 Haratizadeh (2021) [5]将任意一组股票之间的关系建模为一种称为影响网络的图结构，使用图卷积网络算法来预测股票走势，准确率优化后仅有 0.56。知识图谱的运用尚未形成一个完整体系的研究脉络，研究工作还有待改善，其他图片的使用还有待开发。

随着输入数据的多元化，单一的模型已无法兼顾多源异质数据，学者们开始搭建联合模型对数据进行处理。Ho 和 Huang (2021) [6]通过全连接层连接两种 CNN 模型建立集成模型，同时处理蜡烛图和由推

特文本分析得到的情绪评分两类数据，发现联合模型优于单独分析蜡烛图和单独处理推特文本的模型。Li 和 Pan (2020) [7]建立由 LSTM 和 GRU 构成的集成模型，将情感评分和标准普尔 500 指数一起输入进行预测股票价格，发现混合集成深度学习模型在很大程度上优于使用相同数据集的现有最佳预测模型，准确率提高 33.34%。

综上所述，为了克服图片使用的局限性，文章提出利用词云图片数据预测股票走势涨跌。同时为进一步更准确预测股票走势，本文融合历史性交易数据、社交评论数据和词云图片数据三类多源异质数据作为输入数据，并建立基于词云图片的 CNN 和 LSTM 联合模型(wordcloud C-L 模型)来提高预测精度。

2. 基于词云图片的 CNN 和 LSTM 联合模型构建

2.1. LSTM 模型

LSTM 神经网络由 RNN 神经网络演化而来，优化 RNN 易产生梯度爆炸的问题，适用于处理时间序列中长间隔、高延迟事件。LSTM 神经网络原理就是引入了记忆元或简称为单元(cell)，设计其用于记录附加信息。同时引入三个逻辑门用于控制单元，使过去的信息能够被保存，允许其稍后重新进入，从而解决梯度消失问题。三个逻辑门为：输入门(input gate)、输出门(forget gate)、遗忘门(output gate)，用于控制细胞存储状态。LSTM 模型的结构示意图如图 1 所示。

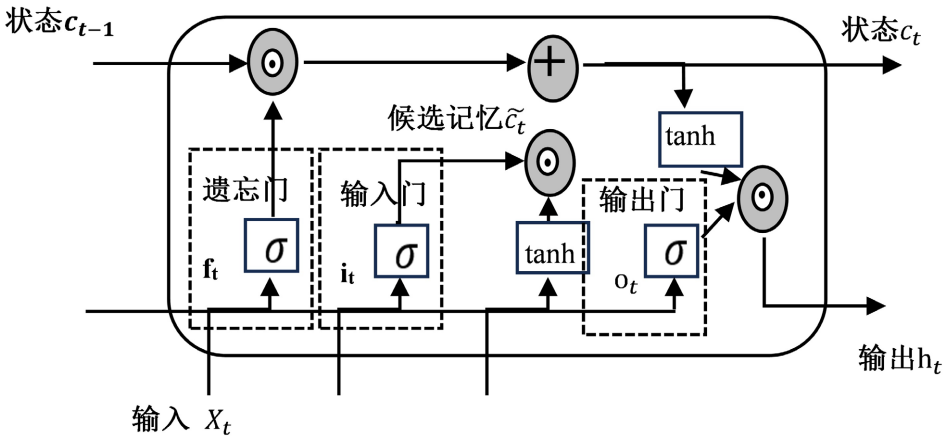

Figure 1. Schematic diagram of LSTM structure
图 1. LSTM 结构示意图

C_t 代表 t 时刻单元状态，X_t 为这一层的输入向量，h_t 为这一层的输出向量，δ 为带激活函数(sigmoid)的全连接层。输入门的作用是将本时间点的输入信息记录到单元状态中。遗忘门的作用是选择性遗忘信息，结合本时间点输入的信息将上一时间点给出的单元状态中不需要的信息移除。输出门决定输出的信息 h_t [8]。

LSTM 的计算公式如下：

输入门：

$$i_t = \sigma\left(W_{xi}X_t + W_{hi}h_{h-1} + b_i\right) \tag{1}$$

i_t 代表输入门，W_{xi}, W_{hi} 表示输入门的权重矩阵，b_i 为输入门的偏置向量。

遗忘门：

$$f_t = \sigma\left(W_{xf}X_t + W_{hf}h_{h-1} + b_f\right) \tag{2}$$

f_t 代表遗忘门，W_{xf}, W_{hf} 表示遗忘门的权重矩阵，b_f 为输入门的偏置。

输出门：

$$o_t = \sigma\left(W_{xo}X_t + W_{ho}h_{h-1} + b_o\right) \tag{3}$$

o_t 代表遗忘门，W_{xo}, W_{ho} 表示遗忘门的权重矩阵，b_o 为输入门的偏置。

$$\tilde{c}_t = \tanh\left(X_t W_{xc} + h_{t-1}W_{hc} + b_c\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{5}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{6}$$

2.2. CNN 模型

CNN 模型是一种带有卷积结构的前馈神经网络，主要应用于图像识别。CNN 神经网络主要结构为卷积层、激活层和池化层，三者可叠加重复使用。核心部分为卷积层，它利用局部连接和权值共享对输入进行卷积操作，提取数据的深层特征[9]。卷积过程如公式(7)所示：

$$C = f\left(X \otimes W + b\right) \tag{7}$$

公式(7)中，W 是卷积核权值向量；X 为输入向量；$\otimes$ 为卷积操作，输入和卷积核权重进行互相关运算，并添加标量偏置(b)后，通过非线性激活函数($f(\ )$)，产生输出特征图(C)。CNN 模型优势在于能够局部感知，每个卷积核仅与输入图像中的一小块区域进行卷积，捕获局部特征，使得模型能够有效地处理图像中的局部结构。同时卷积操作的参数共享，减少了模型参数的数量，降低过拟合的风险，可以较好地处理高维数据。

2.3. 构建情感词典

情感挖掘目前常用且普适性高的方法是使用情感词典，对文本分词后进行情感词匹配，汇总情感词进行评分，得到文本的情感倾向。

中文语句在不同语境中存在一词多义的情况，这给情感倾向的判断带来了很大的困难。目前国内尚未研究出一部可以应用在任何领域的完善的情感词典。因此，进行情感分析的第一步是构建一部适用于自身数据的情感词典。

当前常用的传统情感词典是知网(HowNet)情感分析词典、台湾大学的 NTUSD 中文情感词典以及清华大学李军中文词典。但在本文的研究场景中，有部分词语是股票市场所特有，如：牛市、熊市、跌停、看跌、涨停等。因此本文需要重新建立一部适用于金融领域的情感词典。

首先将前面三种通用词典去除重复的情感词并组合在一起形成基础情感词典，其次加入姚加权金融词典[10]，最后根据常用的股票市场操作词汇表，手动改变一些在金融领域表达情感倾向不同的词语，使其适用于股票市场情感分类的场景，构建一部适合于股票领域的专用情感词典。

2.4. 模型设计

为同时分析图片数据、文本数据和结构化数据，本文构建联合模型 wordcloud C-L。联合模型包含两个分支：一个分支是用于处理词云图片的 CNN 模型，称为 wordcloud CNN。其中包含输入层、卷积层、池化层、Dropout 层；二个分支是处理历史股票交易数据和情感评分数据的 LSTM 模型，其中包含输入层、LSTM 层、Dropout 层。最后将两个分支的输出进行连接，加入 Dense 层，输出股票价格在未来的涨

跌趋势，模型流程图如图 2 所示。

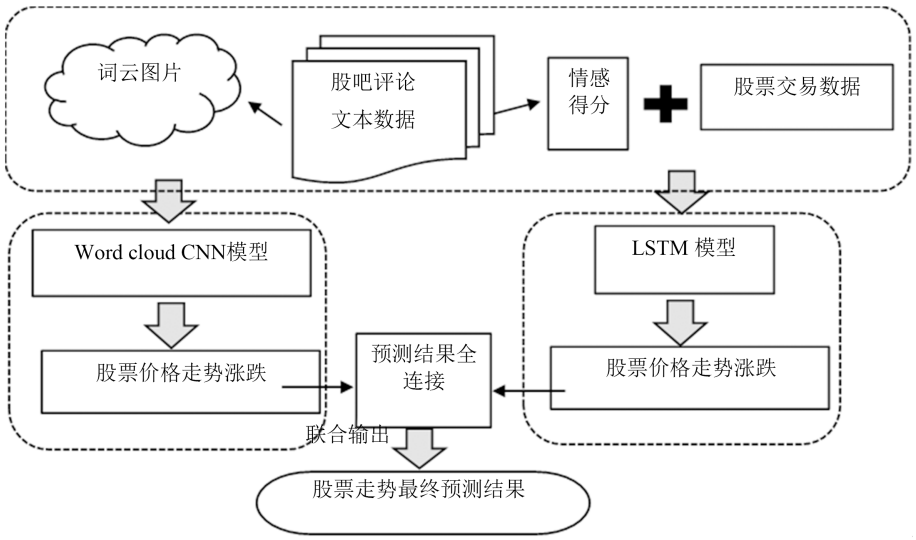

Figure 2. Model flowchart
图 2. 模型流程图

3. 实证分析

3.1. 获取数据及预处理

本文的股票历史交易数据来源于英为财情网，从中获取不同行业的 5 只股票：贵州茅台股票(600519)、沪深 300 指数(000300)、欢瑞世纪股票(000892)、腾讯控股股票(00700)，比亚迪股票(002594)，2021 年 9 月 24 日~2023 年 3 月 28 日的历史交易数据，其中包含四个指标：最高值、最低值、开盘价、收盘价。以收盘价为目标值测定股票价格走势涨跌：将 4 天后收盘价与当天收盘价进行对比，若 4 天后收盘价大于当天收盘价，记为涨，即为 1，否则记为跌，即为 0。同理，建立 6 天后、8 后天股票价格涨跌标签。以贵州茅台股票为例，处理后的历史交易数据如表 1 所示。

Table 1. Historical trading data of Kweichow Moutai stock
表 1. 贵州茅台股票历史交易数据

日期	收盘价	开盘价	最高价	最低价	4 天后股票价格涨跌标签
2021-09-24	1694	1628	1719.98	1628	1
2021-09-27	1855	1750	1863.4	1750	0
2021-09-28	1822.06	1848	1860	1795	0
2021-09-29	1820	1809	1845.01	1785.9	0
2021-09-30	1830	1818.18	1850.22	1803.4	0

使用八爪鱼采集器软件在股吧中获取股民评论数据，根据发帖时间，爬取 2021 年 9 月 24 日~2023 年 3 月 28 日的全部帖子版块的帖子标题。为剔除无关噪声，如：重复贴、广告贴、无关意义的水贴等，需进一步整理爬取的文本。将 10 个字符以内的评论标题看作无意义的水贴，与重复值一并剔除。最终，共获得贵州茅台股票帖子 111,237 个，沪深 300 帖子 5428 个，腾讯控股帖子 34,952 个，欢瑞世纪帖子 29,405 个，比亚迪帖子 102,160 个。以贵州茅台股票和沪深 300 指数为例，评论文本内容如表 2 所示。

Table 2. Stock bar comment text
表 2. 股吧评论文本

时间	贵州茅台股民评论	沪深 300 指数股民评论
2021-09-24	说到底，茅台就是中国股市的灾难……	跟着政策走，跟着概念走，懂一点技术分分析……
2021-09-24	贵州茅台会一直涨。这个价格买了稳赚不赔。	坚持昨天的观点不变！持仓等下步操作指标！耐心点，如
2021-09-24	我想买他拿到明年，可惜我连一手都买不起[哭]	这大盘看着像要完蛋啊！还是先撤吧，安全一
2021-09-24	茅台能到三千元一股，也快到了吧，还要……	牛市来了！！！谁也挡不了！！！
…	…	…
2023-03-28	白酒股年初至今涨幅排名：1、老白干酒：涨 30.58%2、今……	三根阴线不破那根大阳，还是很强的
2023-03-28	业绩增长已是过去震荡慢慢跌吧少也 5 年	唉！在沪指上，只发表选股困难就被禁发，不对等的消息
2023-03-28	未来肯定会上三千，但你得活到老巴的年龄	我不会玩，请问一点几块钱？
2023-03-28	太多存量酒在市场流通，崩盘卖不出是迟早事情	先走高然后下跌下去到尾收盘慢慢涨上来点
…	…	…

将每日的所有帖子内容进行合并，生成每日总评论，将评论的日期与获取的股票历史交易数据日期进行匹配，即只保留同时存在股票历史交易数据和每日总评论的数据。最终得到贵州茅台股票数据 338 个，沪深 300 指数数据 363 个，腾讯控股股票数据 370 个，欢瑞世纪股票数据 364 个，比亚迪股票数据 349 个。

3.2. 特征生成

3.2.1. 词云图片

本文采用 python 中 jieba 模块，对每日总评论进行分词，生成词语列表。对词语列表进行去停用词操作后，统计词频。将词频排名前 200 的词语生成词云图片。词云图片以日期和股票涨跌标签命名，词云图片如图 3 所示。

该图为贵州茅台股票 2021 年 11 月 4 日的总评论生成的词云图片，如果进行 4 天后的股票走势预测，走势涨跌标签应选取预测 4 天后的标签，为“pos”，因此将其命名为：2021-11-04.pos。

Figure 3. Word cloud image
图 3. 词云图片

3.2.2. 情感得分

Python 中 cnsenti 模块的 calculate 方法，能够考虑情感词之前是否有强度副词的修饰和否定情感语义

反转作用，从而精准计算情感信息，得到积极和消极情感得分。本文在 censenti 模块中导入自定义构建的股票专业金融词典，使用 calculate 方法进行情感极性统计，得到评论中积极情感和消极情感词性统计，根据公式(8)计算得到情感评分。

$$sentiment = \frac{pos - neg}{pos + neg} \tag{8}$$

公式(8)中，sentiment 表示情感评分，pos 是评论中积极词性得分，neg 是评论中消极词性得分。sentiment 取值范围为[−1,1]，0~1 为积极评论，越靠近 1 越积极，−1~0 为消极评论，越靠近−1 越消极，靠近 0 为中性评论。以贵州茅台股票评论信息为例，情感评分计算结果如表 3 所示。

Table 3. Kweichow Moutai stock review emotional score calculation results
表 3. 贵州茅台股票评论情感得分计算结果

序号	pos	neg	words	number
1	2209.50	1997.00	1452.00	0.05
2	7769.00	3637.50	5726.00	0.36
3	522.50	451.00	379.00	0.07
4	7282.00	5158.50	4847.00	0.17
5	4848.00	2532.00	3006.00	0.31

其中，pos、neg 为评论中积极情感、消极情感得分，words 是对评论进行分词处理得到的词语总数，number 为根据公式(8)得到的情感评分。

3.3. 评价指标

根据样本真实情感与模型预测情感可将测试数据的结果划分为正确正例(TP)、错误正例(FP)、正确负例(TN)、错误负例(FN)这 4 类，本文采用准确率(Accuracy)对模型结果进行评估。准确率计算如公式(9)所示：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

3.4. 结果分析

为表明联合模型处理多源异质数据的优势，本文采用实证研究将每分支单一模型与联合模型进行比对实验。即在“历史交易数据 + 情感评分”数据集进行 LSTM 模型测试，在“词云图片”数据集进行 wordcloud CNN 模型测试，在“词云图片 + 历史交易数据 + 情感评分”数据集进行 wordcloud C-L 模型测试，每个数据集都包括 5 只股票，测试模型的通用性。贵州茅台股票是白酒行业具有代表性的股票，沪深 300 指数由沪深市场中规模大、流动性好的最具代表性的 300 只证券组成，可反映沪深市场上市公司证券的整体表现。两只股票具有代表性，故以两者为例，输出迭代 30 次模型对比的结果，如图 4、图 5 所示。

图 5 为沪深 300 指数在三个模型中迭代 30 次的结果，图例中的数字尾号代表模型预测的日期为 4 天、6 天或者 8 天，图 6 为贵州茅台股票在三个模型中迭代 30 次的结果。为增强实证结果的可靠性，本文输出的准确率为模型迭代 30 次的平均准确率，输出结果如表 4 所示。

通过比对实验得到以下结论：

1) 图 4、图 5 表明单一模型结果存在波动，准确率在 0.4~0.6 之间跌宕，而 wordcloud C-L 模型的曲线平稳，准确率稳定在 0.6~0.7 之间，效果较单一模型有所提升。表 4 显示 5 只股票每个模型的平均准确

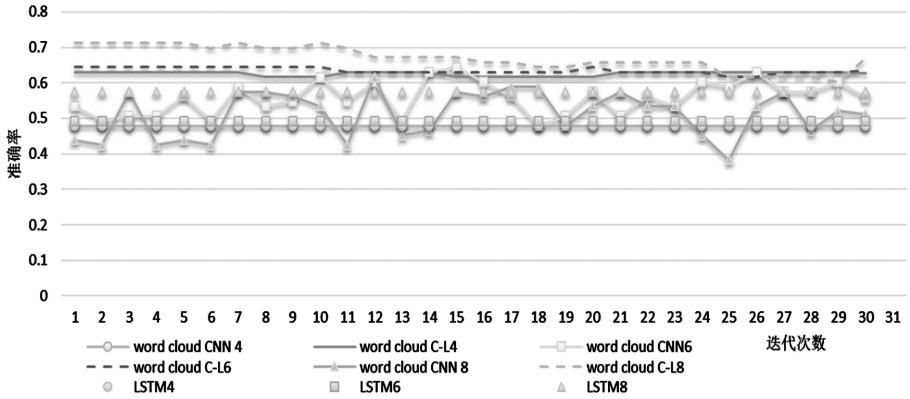

Figure 4. The Shanghai and Shenzhen 300 model results for 30 interactions
图 4. 沪深 300 迭代 30 次模型结果

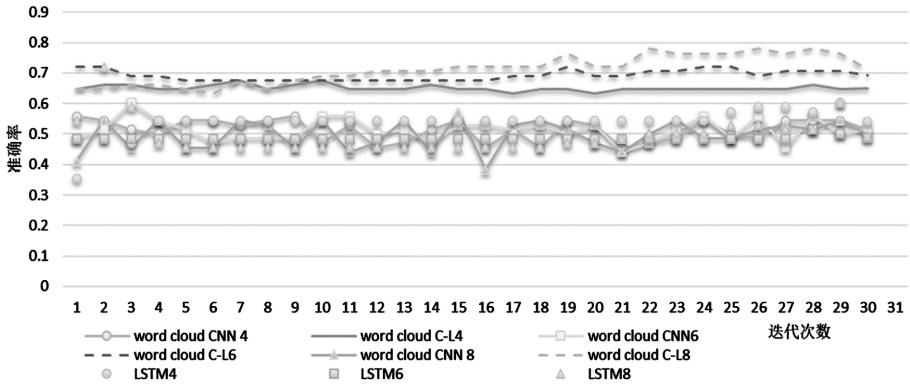

Figure 5. Kweichow Moutai model results for 30 interactions
图 5. 贵州茅台迭代 30 次模型结果

Table 4. Model accuracy output results
表 4. 模型准确率输出结果

股票名称	时间	LSTM	wordcloud CNN	wordcloud C-L
贵州茅台(600519)	4 天	0.54	0.51	**0.65**
	6 天	0.49	0.51	**0.69**
	8 天	0.49	0.50	**0.71**
腾讯控股(00700)	4 天	0.45	0.49	**0.52**
	6 天	0.36	0.42	**0.47**
	8 天	0.35	0.45	**0.51**
比亚迪(002594)	4 天	0.61	0.52	**0.65**
	6 天	0.63	0.53	**0.64**
	8 天	0.62	0.53	**0.64**
欢瑞世纪(000892)	4 天	0.47	0.51	**0.53**
	6 天	0.49	0.54	**0.55**
	8 天	0.50	0.51	**0.58**
沪深 300 (000300)	4 天	0.48	0.48	**0.63**
	6 天	0.49	0.56	**0.63**
	8 天	0.58	0.51	**0.67**

率，5 只股票代表不同行业，数据本身具有差异，故而在单一模型中的表现有所不同，可联合模型均能较之改善。例如：比亚迪股票在 LSTM 模型效果良好，wordcloud C-L 联合模型增幅 3%~6%；腾讯控股股票在 LSTM 模型效果不佳，wordcloud C-L 联合模型增幅显著，表明联合模型的优势。对词云图片进行卷积处理的 wordcloud CNN 模型，在欢瑞世纪、腾讯控股、贵州茅台股票的测试结果，优于处理“历史交易数据 + 情感得分”的 LSTM 模型，表明词云图片的可用性。

2) 为验证预测的时效性，将三个模型同一时间预测的准确率进行平均，得到每只股票对应日期的平均准确率，如图 6 所示。

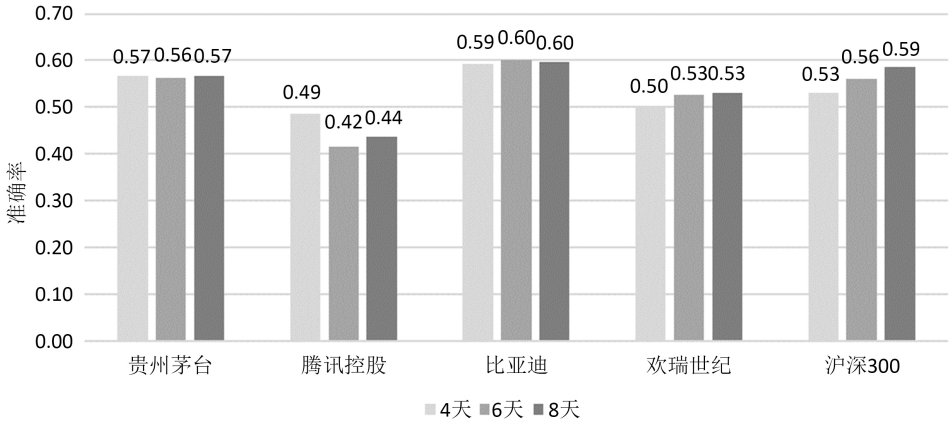

Figure 6. Prediction results of different time models for stocks
图 6. 股票不同时间模型预测结果

除腾讯控股以外四只股票，从 4 天到 8 天，预测精度逐渐提高，沪深 300 八天后准确率较四天后提升 11%，表明预测模型在较长时间段内的表现优于较短时间段。这可能是因为评论文本对于投资者情绪和行为的影响需要时间来发酵。因此，当采用影响情绪数据进行预测时，应考虑较长时间的标签，有助于提高预测的准确性。

4. 结论

本文针对基于图片的股票预测模型，提出基于词云图片的 wordcloud C-L 联合模型。将股吧评论文本转为词云图片并运用 wordcloud CNN 进行模型处理，同时重新编写适用于股票的情感字典，从而得到评论文本的情感得分，与历史交易数据一起运用 LSTM 模型处理。将两个模型联合，得到 wordcloud C-L 联合模型，共同预测股票走势涨跌。

实证研究表明词云图片可以用于股票走势预测，本文提出的联合模型能够兼顾多源异质数据，模型效果不但优于两个单一模型，而且具有稳定性，还在不同行业股票的预测中准确率均具有良好表现。在未来的工作中，考虑到利用多源异质的联合模型，可以研究多源数据的权重选择，加入注意力机制等方法，进一步提高模型的预测精度。预测模型涉及到情感分析时，可以选择时间跨度长的天数，例如预测 10 天、30、100 天等，探索预测模型最适合的时间。

参考文献

[1] 康瑞雪, 牛保宁, 李显, 等. 融合多源数据输入具有自注意力机制的 LSTM 价格预测[J]. 小型微型计算机系统, 2023, 44(2): 326-333.

[2] Maqbool, J., Aggarwal, P., Kaur, R., *et al.* (2023) Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. *Procedia Computer Science*, **218**, 1067-1078.

https://doi.org/10.1016/j.procs.2023.01.086

[3] Liu, S., Zhang, X.D., Wang, Y., *et al*. (2020) Recurrent Convolutional Neural Kernel Model for Stock Price Movement Prediction. *PLOS ONE*, **15**, e0234206. https://doi.org/10.1371/journal.pone.0234206

[4] Hung, C.C. and Chen, Y.J. (2021) DPP: Deep Predictor for Price Movement from Candlestick Charts. *PLOS ONE*, **16**, e0252404. https://doi.org/10.1371/journal.pone.0252404

[5] Jafari, A. and Haratizadeh, S.G. (2022) CNET: Graph-Based Prediction of Stock Price Movement Using Graph Convolutional Network. *Engineering Applications of Artificial Intelligence*, **116**, Article ID: 105452. https://doi.org/10.1016/j.engappai.2022.105452

[6] Ho, T.T. and Huang, Y. (2021) Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. *Sensors*, **21**, Article No. 7957. https://doi.org/10.3390/s21237957

[7] Li, Y. and Pan, Y. (2022) A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News. *International Journal of Data Science and Analytics*, **13**, 139-149. https://doi.org/10.1007/s41060-021-00279-9

[8] 刘月娟, 王武. 基于多特征融合的股票走势预测研究[J]. 云南民族大学学报(自然科学版), 2022, 31(2): 227-234.

[9] 方义秋, 卢壮, 葛君伟. 联合 RMSE 损失 LSTM-CNN 模型的股价预测[J]. 计算机工程与应用, 2022, 58(9): 294-302.

[10] 姚加权. 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.