東南大學

本科毕业设计（论文）报告

一种基于深度学习的股票预测算法的设计与实现

学　　号: 58121232

姓　　名: 胡鑫月

学　　院: 人工智能学院

专　　业: 人工智能

指导教师: 吕建华

起止日期: 2024.12-2025.5

2025 年 6 月 13 日

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：卞刚鑫 日期：2015年6月13日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：卞刚鑫 导师签名：

日期：2015年6月13日 日期：2015年6月13日

东南大学本科毕业论文（设计）AI 工具使用情况说明表

课题名称	一种基于深度学习的股票预测算法的设计与实现		
学　　号	58121232	姓　　名	胡鑫月
是否使用生成式人工智能	■是　　□否		
工具、版本号	使用范围	使用过程	章节（页码）
选择是：DeepSeek-RI	■文本生成及内容修改 □数据、图表分析、代码调试 □其他：请具体说明	在对实验结果进行分析时，输入"情绪因子对不同股票的影响"，使用 AI 工具生成参考分析视角.	第四章 4.4.3（P45）
学生诚信申明	本人郑重声明，上述关于生成式人工智能使用情况的陈述真实无误，已对使用此类技术的所有细节进行了全面且诚实的报告。本人深知学术诚信的重要性，如有任何隐瞒或虚假之处，愿承担学术不端行为带来的相关惩处。 学生签名：胡鑫月 2025 年 6 月 13 日		
指导教师意见	意见： 情况属实 指导教师签名：[illegible] 2025 年 6 月 13 日		

摘　　要

股票市场的价格波动，对于经济发展以及个人投资决策而言有着关键的影响，依据股票市场的历史数据、舆论所呈现出的情绪、宏观经济政策等信息来对未来股价加以预测，这对于投资者识别潜在的投资机会以及风险，优化资产配置策略来说有着重要意义。

传统的数学统计模型在股票价格预测方面表现不太理想，主要是受到线性假设、对噪声比较敏感以及依赖历史数据等问题的限制。深度学习模型依靠其强大的非线性建模能力以及多模态数据融合能力，在一定程度上提高了股票价格预测的准确性，但依然面临着数据特征提取不全面、模型解释性欠佳、市场情绪的考量以及多模态特征融合等挑战。

本文基于基本深度学习框架，提出了一种基于多层次特征融合的股票预测模型 CNN-LSTM-Attention，可有效提取局部时序特征，建模长期依赖关系，聚焦关键信息，从而实现数据特征的多层次提取，服务于股票预测任务。本文从 Tushare 平台获取中国证券市场 CSI-300 指数的成分股数据，将本文所提出模型与传统深度学习模型对比，证实了 CNN-LSTM-Attention 融合模型在股票预测任务上的优越性能。

为深入探索市场情绪对于股票趋势的影响，本文在 CNN-LSTM-Attention 模型的基础上提出了一种基于市场情绪融合的股票预测模型，在分别对股票数值数据与情绪数据进行特征提取之后，基于交叉注意力机制进行特征融合，用于股票预测任务。本文从东方财富网股吧收集了三只股票近六年的新闻和评论数据，根据金融情感词典计算情绪指数，实验结果表明，融合市场情绪的预测模型在股票预测中表现更为出色。

关键词：股票价格预测，深度学习，多模态数据融合，情感分析

ABSTRACT

Stock market price fluctuations play a crucial role in economic development and individual investment decisions. Predicting future stock prices based on historical market data, public sentiment, and macroeconomic policies is essential for investors to identify potential opportunities and risks, and optimize asset allocation strategies. Traditional statistical models often underperform in stock price prediction due to its linear assumptions, sensitivity to noise, and overreliance on historical data. While deep learning models, with their superior nonlinear modeling and multimodal data fusion capabilities, have improved prediction accuracy, challenges remain, including incomplete feature extraction, poor interpretability, and inefficient multimodal fusion.

This thesis proposes a multi-level feature fusion stock prediction model CNN-LSTM-Attention based on the basic deep learning framework, which can effectively extract local temporal features, model long-term dependencies, focus on key information, and achieve multi-level feature extraction of data to serve stock prediction tasks. Experiments on CSI-300 index component stocks from the Tushare platform demonstrate the model's superiority over traditional deep learning approaches.

To further explore the impact of market sentiment, this thesis proposes a stock prediction model that integrates market sentiment. On the basis of CNN-LSTM-Attention, multi-level feature extraction is performed on both numerical and emotional data, followed by feature fusion using a cross-attention mechanism for stock prediction tasks. With the sentiment index calculated based on the Financial Sentiment Dictionary, experiments on news and comments collected from East Money show that the prediction model integrating market sentiment performs better in stock prediction.

KEY WORDS: Stock price prediction, Deep learning, Multimodal data fusion, Sentiment analysis.

目　录

第一章 绪论

1.1 课题背景和意义

中国股票市场起步晚于西方国家，但在改革开放的进程当中实现了跨越式的发展。截至 2024 年末，沪深京三大交易所的市值突破了 86.01 万亿元，其中在 2024 年新上市的 100 只股票贡献了 8595 亿元。股票市场作为现代经济体系的核心枢纽，对投资者来讲，可提供资产增值的途径，同时还可分散投资风险；对企业来讲，可为企业提供直接的融资平台，帮助企业进行科技创新以及产业升级；对国家来讲，可借助市场供需机制实现资产价值发现，引导资本朝着高成长性领域流动，同时也能反映国民经济的健康状况。

我国股票市场呈现出波动性较强、对政策敏感、投机性较高、市场结构复杂等特点，这些特点增加了投资风险以及市场预测的难度。故建立科学、准确的价格预测系统是非常关键的，它可帮助投资者应对高波动性、降低投资风险，对于监管机构预判系统性风险、优化资产配置也有益处。股票的长期预测和短期预测在目标、方法以及应用方面存在差异。长期预测一般关注宏观经济因素、行业趋势以及公司基本面，时间跨度可能长达数年，可为投资者提供战略性决策依据；相对而言，短期预测更注重市场情绪、技术指标以及高频数据，时间跨度一般为几天到几周，适合短线交易者捕捉市场波动。本文主要聚焦于短期预测，分析市场动态和情绪变化，为投资者提供更精准的交易信号。

早期的股票预测研究受限于计算机技术，多采用统计和经济计量模型拟合股价，如 ARIMA（自回归积分滑动平均模型）和 GARCH（广义自回归条件异方差模型）等，但传统的数学模型难以考虑股票市场的多种影响因素，具有很大的局限性。随着机器学习的诞生，研究者开始对股票数据进行非线性建模，机器学习通过模拟生物神经网络，能够挖掘数据的非线性变化。数据质量和数量的提升以及计算机性能的提高使得深度学习渐渐成为研究热点，深度学习模型能够自动学习数据特征，相比浅层机器学习，它可以直接处理时间序列数据，如 LSTM（长短期记忆网络）可以捕捉时间序列中的长期依赖关系，除此之外，近几年的 Transformer 模型及其变体模型利用注意力机制在时间序列预测任务上表现也很出色。

当前用于股票预测的深度学习模型大多集中在数值特征方面，而忽视了市场

情绪对股票预测的影响，市场情绪是影响股价短期波动的关键因素之一。研究发现[1]，投资者情绪得分与股票收益率之间存在正相关关系，投资者情绪提升会使股票收益率增加。情绪信息的获取和量化仍面临挑战：情绪数据来源多样且噪声大，从中提取有效信息是一个难题，并且情绪与股价之间的关系复杂，可能受到多种因素干扰，比如市场环境、投资者结构和信息传播速度等。随着自然语言处理技术和情感分析算法的进步，情绪信息在股票预测中的应用前景广阔。

综上所述，本文将运用深度学习的方法，整合 CNN、LSTM 以及 Attention 机制来开展股票价格的短期预测工作，同时关注市场情绪，融合市场情绪信息，以此提升深度学习模型在股票预测方面的准确率。

1.2 国内外研究现状

1.2.1 传统统计模型

早期关于股票预测的研究受计算机技术以及算法的限制，大多采用传统统计与计量模型，如 ARMA（自回归移动平均模型）、ARIMA（自回归积分滑动平均模型）以及 GARCH（广义自回归条件异方差模型）等。ARIMA 运用差分和滞后项处理平稳时间序列，Banerjee 等人[2]运用 ARIMA 对孟买证券交易所敏感指数月度收盘数据构建模型并预测未来指数，验证了 ARIMA 模型在股票市场预测的有效性。 GARCH 用于对波动率的聚类效应进行建模，Maqsood 等人 [3] 应用 GARCH 模型对肯尼亚内罗毕证券交易所的股票市场波动性做了建模和分析，找寻最能有效捕捉其波动性和杠杆效应的模型。这些方法存在明显的不足：它们假定股价变动是线性的，很难捕捉复杂的非线性关系；这些数学模型对突发事件的适应性欠佳，缺少对市场情绪、宏观政策等因素的考量，无法有效处理市场中的异常波动，而且，它们对高频数据的处理能力有限，难以契合现代金融市场对实时预测的需求。

1.2.2 机器学习模型

随着互联网技术的普及以及机器学习算法的兴起，股票预测研究逐渐转向非线性数据建模。机器学习方法依靠捕捉数据中的复杂模式，提高了预测精度。支持向量机（SVM）以核函数把数据映射到高维空间，可有效处理非线性关系， Lin 等人[4]对股票数据进行特征选择并赋予权重，使用一种准线性核函数的支持向量机进行股票预测，在台湾股市数据集上表现良好，有效避免了过拟合问题；随机

森林依靠构建多个决策树并进行集成学习，适用于具有复杂模式的时间序列数据，Ren Zi 等人[5]提出了一种基于蚁群算法优化的加权随机森林模型，在平安银行、贵州茅台等四只股票数据上，其预测误差明显低于普通随机森林、梯度提升决策树和决策树模型。

近年来，随着股票数据量大幅增加，深度学习模型因其出色的特征学习能力，被广泛应用于股票预测任务。CNN（卷积神经网络）依靠卷积核提取时间序列中的局部特征，可捕捉数据中的短期模式和空间依赖性，Rachna Somkunwar 等人[6]采用 CNN 模型结合多元线性回归来预测股票价格的未来走势，并在 NIFTY 50 指数（Nifty Fifty Index，由印度国家证券交易所编制，包括 50 只在印度国家证券交易所上市的大型、流动性强的股票，涵盖了印度经济各主要行业）数据集上获得出色成效，模型准确率达到了 97.67%；LSTM（长短期记忆网络）在股票预测中表现得极为突出，它可以捕捉时间序列里的长期依赖关系，处理股价数据当中的趋势以及周期性变化，Bathla 等人[7]使用 LSTM 模型来预测 2020 年的股票市场波动，证明了其在高波动性情况下是有效得；GRU（门控循环单元）引入重置门和更新门，缓解了 RNN 的梯度消失问题，Umang Gupta[8]提出了一个基于 GRU 的 StockNet 模型，该模型借助注入模块（Injection Module）进行数据增强，注入不同特征，用调查模块进行股票预测，运用启发式方法选择输出神经元的数量，此模型在印度国家证券交易所 CNX-Nift 指数的历史开盘价数据方面表现良好，RMSE、MAE 和 MAPE 有所降低；Chirag Choudhury 等人[9]提出一种 LSTM 与 GRU 的混合模型，并且用遗传算法优化模型超参数，在 SPDR SP 500 ETF (SPY)数据集上表现出色。然而这些模型对数据质量和数量的依赖程度较高，容易出现过拟合风险，需要依靠数据增强或者结合其他算法，而且模型的可解释性较差。

2017 年提出的注意力机制[10]，在自然语言处理、计算机视觉和时间序列分析等领域取得了广泛应用。注意力机制可以捕捉长期依赖关系，解决传统深度学习模型的梯度消失问题，进行动态权重分配，自动聚焦于时间序列中的重要部分，并行计算也大大提高了计算效率。Wang 等人[11]使用 Transformer 模型在包括中国沪深 300 指数（CSI 300）、美国标准普尔 500 指数（S&P 500）、日本日经 225 指数（Nikkei 225）和香港恒生指数（Hang Seng Index）在内的全球主要股票市

场指数上实施了回测实验，结果表明 Transformer 在预测准确性上显著优于其他经典方法，并且能够为投资者带来超额收益。还有一系列基于 transformer 的变体模型，通过改进模块使其更适应时间序列预测任务。Informer[12]通过概率稀疏(ProbSparse)自注意力机制和蒸馏操作来解决传统 Transformer 在时间序列预测中的不足。Reformer[13]它使用局部敏感哈希（LSH）来代替传统的点积注意力机制，降低计算复杂度，运用可逆残差层（Reversible Residual Layers）代替标准的残差链接降低内存使用，在合成任务、文本任务和图像任务上实现了更高的运算速度和内存效率，可以考虑将它运用于股票预测任务。FlashAttention[14]采用了一种 IO 感知的注意力算法，通过减少 GPU 高带宽内存（HBM）和片上 SRAM 之间的内存读写次数，显著降低了注意力计算的内存访问开销，实现更少的内存访问从而提高运行速度。Flowformer[15]引入了一种基于流网络的注意力机制，将注意力机制重新解释为信息从源（值）到汇（结果）的流动，在保持模型通用性的同时生成信息丰富的注意力分布，在长序列建模、时间序列分析等多个领域的基准测试中表现出色。iTransformer[16]将 Transformer 的注意力机制和前馈网络应用于倒置的维度，通过注意力机制捕捉多变量之间的相关性，，在电力消耗、交通流量等数据集上进行多元变量预测时表现出了出色性能。

将各类深度学习模型的优势相结合形成融合架构，这对提升股票预测精度是有帮助的。把 GRU 与注意力机制[17]结合起来，模型可有效地捕捉数据里的时间依赖性，同时依靠注意力机制动态关注输入序列中最具相关性的部分，以此提高预测的准确性。融合 Time2vec 和 Transformer[18]，模型借助时序编码来建模周期性规律，利用 Transformer 层提取高阶时空特征，捕捉市场数据中固有的时空依赖。

尽管这些深度学习模型在股票预测方面取得了一定进展，但它们主要依赖历史价格、交易量等结构化数据，却忽略了市场情绪等非结构化数据对股票价格的影响。实际上，市场情绪作为投资者心理的一种反映，大多情况下对股票价格的短期波动有着关键作用。

1.2.3 情感分析

市场情绪是指投资者对市场整体或特定资产的集体心理状态，对股票价格的短期波动具有显著影响，尤其是在市场不确定性较高时，情绪因素往往成为驱动

价格变化的关键因素。近年来，随着自然语言处理（NLP）技术的进步，研究者开始将市场情绪纳入股票预测模型中。Dong 等人[19]提出了一个综合型的金融时间序列数据集，包括数值信息和情感信息，情感信息由 chatGPT 对总结过后的金融新闻数据进行评分获得，验证了数据集的规模和质量能够显著提高市场预测的准确性。通过分析新闻、社交媒体、财报电话会议等文本数据，可以提取出市场情绪指标，并将其作为预测模型的输入。通过 FinBERT[20]和循环神经网络（RNN）进行股票价格预测，在苹果、微软、帕森斯公司和 PBF 能源四家公司数据集上验证了加入新闻情感得分后，模型的预测性能显著提高且小盘股对新闻情感的反应更为敏感。S_I_LSTM[21]结合了股票历史数据、论坛帖子和金融新闻等多种数据源，使用卷积神经网络（CNN）进行情感倾向分析，计算投资者的情感指数，利用 LSTM 网络处理时间序列数据，比单一数据源的预测方法更准确。结合 LSTM 与注意力机制[22]，能够捕捉时序数据中的长期依赖关系并过滤股价时序噪声，进一步融合宏观经济指标、金融新闻情感分析及历史股价数据，增强了模型对市场多维情境的感知能力。采用词典法和 BERT-BiLSTM 进行融合情感特征和股票交易特征的股指预测[23]，证实了情感特征可以提高预测准确率。结合 GRU 与轻量级大模型 ALBERT 进行股票预测[24]，能够充分利用客观特征以及主观情感，提高了股票预测的准确率。针对文本上下文语义信息引起的噪声扩散，研究者提出了融合全局注意力和局部注意力机制分层的深度学习框架[25]，获取更丰富的语义表征，提高股票预测的准确率。

1.3 本文研究内容

本文主要研究基于深度学习的股票价格预测模型，传统深度学习模型 GRU、LSTM 等在处理时间序列任务时容易出现过拟合情况，而基于 transformer 的模型存在推理速度慢、内存体量大的问题，故本文提出一种基于多层次特征融合的股票预测模型，整合 CNN、LSTM 与 Attention 机制，CNN 部分擅长从局部时间窗口里提取时间序列数据的特征。依靠卷积层，模型可捕捉输入数据中的局部模式和趋势；LSTM 网络能够有效地处理和建模序列数据中的长期依赖关系，借助其独特的门控机制（输入门、遗忘门和输出门）来控制信息的流动，避免长期依赖问题中的梯度消失或梯度爆炸问题；Attention 机制让模型在处理序列时可关注更关键的时间步。这利于模型识别和强调对预测结果影响最大的输入部分，提

高预测的精准度。

为融合市场情绪，在 CNN-LSTM-Attention 多层次特征融合模型的基础上，设置 CNN-LSTM 情绪特征提取模块，多层次提取特征，并通过交叉注意力机制融合数值特征与情绪特征，实现两方面特征的关联。实验结果表明，基于市场情绪融合的模型在股票预测任务中性能更优。

1.4 本文的组织架构

根据本文研究内容，分为五个章节，内容安排如下：

第一章：绪论。本章主要探究股票价格预测的研究背景及重要性，并对国内外研究者对股票价格预测模型的研究现状以及股票市场的情感分析进行整理和综述，最后对本文主要研究内容及组织架构进行说明。

第二章：相关技术基础。本章主要介绍了卷积神经网络、长短期记忆网络以及注意力机制等深度学习方法，并阐述了基于金融情感词典和基于 SnowNLP 库的情感分析方法，以用于股票预测模型的构建。

第三章：基于多层次特征融合的股票预测模型。本章首先介绍多层次特征融合模型 CNN-LSTM-Attention 的模型框架与实现，由 CNN、LSTM、Attention 机制分别提取各层次特征；其次阐述实验设置包括数据集的获取与预处理以及模型参数设置；最后介绍实验所用评价指标，并进行实验结果展示与分析。

第四章：基于市场情绪融合的股票预测模型。本章在第三章多层次特征融合模型 CNN-LSTM-Attention 的基础上引入交叉注意力机制，先通过 CNN-LSTM 分别提取数值与情绪特征，再运用交叉注意力机制融合；接着，本章介绍了情绪文本数据的获取、预处理以及转为情绪指数的过程，阐述该部分的实验设置；最后从多个角度对实验结果进行了分析。

第五章：总结与展望。本章梳理总结本文的研究内容和结论，表明本文研究成果对于提高股票价格预测的准确性具有重要意义；同时，阐明工作中的不足之处，并提出相应的改进方向，以在复杂的市场变化中进一步提高股票预测性能。

第二章 相关技术基础

2.1 深度学习模型

2.1.1 卷积神经网络（CNN）

卷积神经网络[25]最初是在 1998 年被提出的，它定义了 CNN 的卷积层、池化层、全连接层的基本结构。卷积神经网络在图像处理领域表现优异，这是由于它借助卷积运算从图像的局部区域提取特征，并且利用参数共享和层次化表示高效捕捉不同部位的相似特征。而时间序列也可被看作一个二维向量，序列的时间窗口和特征类似于二维图像的高度和宽度。在时间序列任务中，其模型原理如图 2-1 所示。

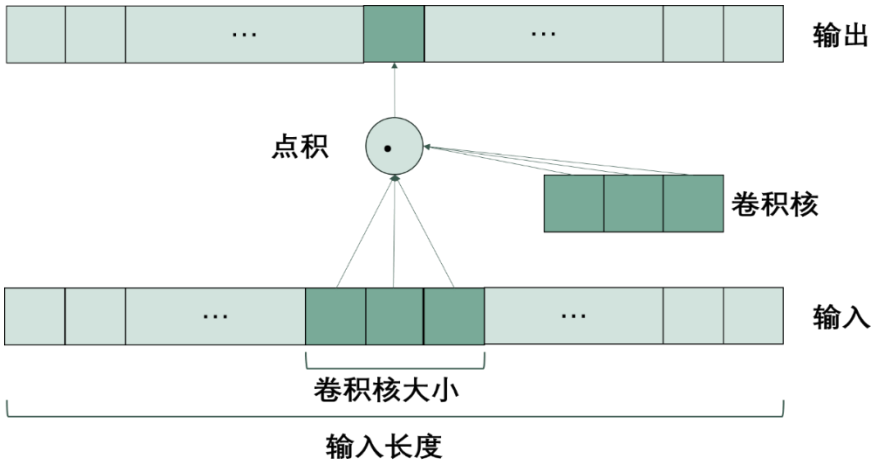

图 2-1 CNN 处理时间序列原理图

完整的一维卷积网络包含输入层、一维卷积层、池化层、展平层、全连接层，如图 2-2 所示。

（1）输入层：用于接收时间序列数据，形状为（T, F），其中 T 是时间窗口长度，F 是每个时间步的特征数量。

（2）一维卷积层：使用一维卷积核在时间维度上滑动，提取时间序列的局部特征，以参数共享机制减少模型参数量。卷积核大小为 k，表示每次观察 k 个时间步，数量为 N，表示提取 N 种不同的特征。在卷积操作后应用非线性激活函数如 ReLU、tanh，过滤掉不相关的特征信息。

$$ReLU(x) = max(0, x) \tag{2.1}$$

ReLU 函数会把所有的负值都设置为 0，而正值则保持不变，如公式（2.1）

所示，这种特点可有效地缓解梯度消失的问题。

$$Tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.2}$$

Tanh 函数会将输入映射到 (-1, 1) 的区间内，其输出是关于原点对称的，如公式（2.2）所示，但可能存在梯度消失的问题。

（3）池化层：一般分为平均池化和最大池化两种。平均池化会提取局部时间窗口内的平均值，以此来反映整体的趋势，在时间序列预测或者平滑任务中有着较好的表现；最大池化则会提取局部时间窗口内的最大值，比较适合捕捉序列的峰值以及拐点，在时间序列分类或者异常检测任务中表现不错。

（4）展平层：会将多维特征展平为一维向量，以方便全连接层进行处理。

（5）全连接层：会综合所有提取出来的特征，去学习特征之间的非线性关系，然后映射到目标输出。

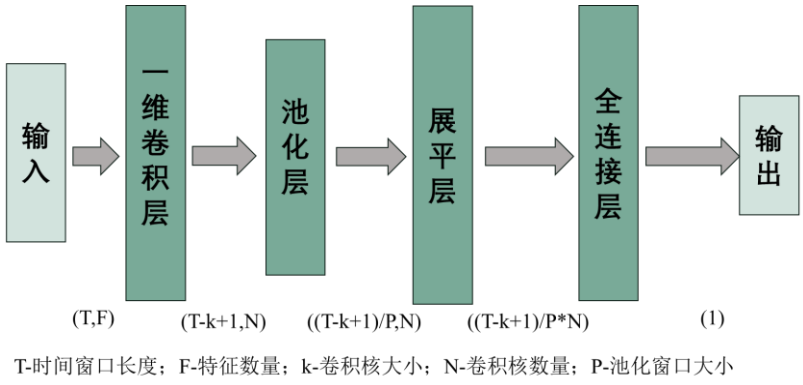

图 2-2 CNN 网络

在时间序列预测中，一维卷积核会沿着时间维度进行滑动，提取局部时间窗口内的特征模式（如短期趋势或周期性波动）。借助堆叠多个卷积和池化层，网络可逐步提取更高层次的时间特征，并且降低序列的长度。全连接层会把这些特征映射到目标输出，比如未来某一时刻的值。这种方法的优势在于 可高效地捕捉时间序列当中的局部模式，同时运用参数共享来减少计算的复杂度，适用于股票价格、天气预测等任务。

2.1.2 长短期记忆网络（LSTM）

在处理时间序列时，RNN 在反向传播梯度的过程中，梯度会随着时间步的增加而呈现指数级衰减，这就导致网络很难学习长期依赖关系，而 LSTM 采用了

门控机制，使得网络可有选择性地记住或者忘记信息，有效捕捉时间序列当中的长期依赖关系。RNN 和 LSTM 的主要区别在于，RNN 只有一个简单的隐藏状态h_t，是借助循环连接传递信息的，而 LSTM 引入了细胞状态和门控机制，其中包括遗忘门、输入门、输出门，可更加精细地控制信息的流动。

LSTM 的核心结构包括以下几个部分：

（1）细胞状态：细胞状态C_t是 LSTM 的核心，用于长期保存信息，它通过遗忘门和输入门的控制，选择性地更新或丢弃信息。候选细胞状态$\tilde{C}_t$表示可能添加到细胞状态C_t中的新信息。

（2）遗忘门：决定哪些信息需要从细胞状态中丢弃。

（3）输入门：决定哪些信息需要保存到细胞状态中。

（4）输出门：决定那些信息需要从细胞状态中输出。

LSTM 的结构如图 2-3 所示：

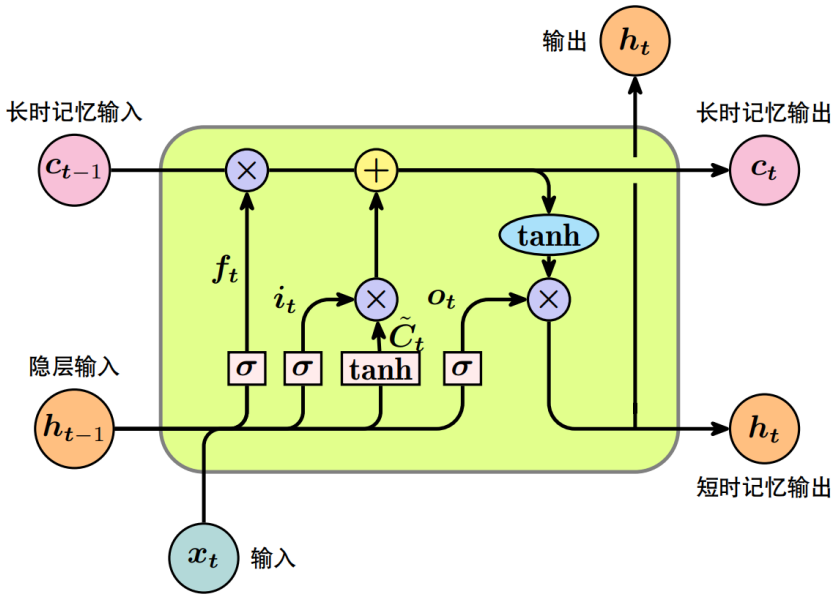

图 2-3 LSTM 结构图

数据在 LSTM 中的传播过程如下：

（1）输入数据：当前时间步的输入x_t和上一个时间步的隐藏状态h_{t-1}被输入到 LSTM 中。

（2）遗忘门计算：通过公式（2.3）计算遗忘门的输出f_t，决定哪些信息需要从细胞状态C_{t-1}中丢弃，当前时间步输入x_t与上一时间步隐藏状态h_{t-1}拼接后

输入激活函数，输出f_t。

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.3}$$

其中W_f是遗忘门的权重矩阵，b_f是遗忘门偏置，σ是 sigmoid 激活函数。

（3）输入门计算：公式（2.4）计算输入门的输出i_t，公式（2.5）计算候选细胞状态$\widetilde{C_t}$，决定哪些新信息需要保存到细胞状态中，当前时间步输入x_t与上一时间步隐藏状态h_{t-1}拼接后输入激活函数，得到i_t，以此控制候选细胞状态$\tilde{C}_t$中需要存储的信息。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.4}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.5}$$

（4）更新细胞状态：结合遗忘门和输入门的结果将旧的细胞状态更新为新的细胞状态，如公式（2.6）所示，通过遗忘门的输出f_t来选择遗忘旧细胞状态C_{t-1}的一部分，通过输入门的输出i_t来选择记忆候选细胞状态$\tilde{C}_t$的一部分，得到新的细胞状态C_t：

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \tag{2.6}$$

（5）输出门计算：计算输出门输出o_t，如公式（2.7）所示，决定哪些信息需要从细胞状态中输出：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.7}$$

（6）计算隐藏状态：计算当前时间步的隐藏状态h_t，如公式（2.8）所示：

$$h_t = o_t \bigodot tanh(C_t) \tag{2.8}$$

（7）输出结果：隐藏状态h_t可以作为当前时间步的输出，也可以传递到下一个时间步。

2.1.3 注意力（Attention）机制

注意力机制[10]通过计算输入序列中不同部分的权重，动态聚焦于关键信息，它的核心是计算查询（Query）与键（Key）的相似度，生成权重分布，再对值（Value）加权求和。注意力机制可以捕捉长距离依赖，提高模型对关键信息的关注，提升序列建模能力，广泛应用于机器翻译、时间序列预测等任务。

在注意力机制中，查询表示当前需要关注的内容，键表示输入序列的特征，用于计算与查询的相似度，值表示输入序列的实际值，用于加权求和生成输出。

在处理时间序列数据时，查询表示目标时间步的特征表示，键表示历史时间步的特征表示，值同样是历史时间步的特征表示，但与键的区别在于，值用于加权求和生成输出。

通过线性变换生成查询、键、值，如公式（2.9）所示：

$$Q = XW_Q, K = XW_K, V = XW_V \tag{2.9}$$

其中，X 是输入序列，W_Q、W_K、W_V是科学系的权重矩阵，用于将输入映射到查询、键和值空间。

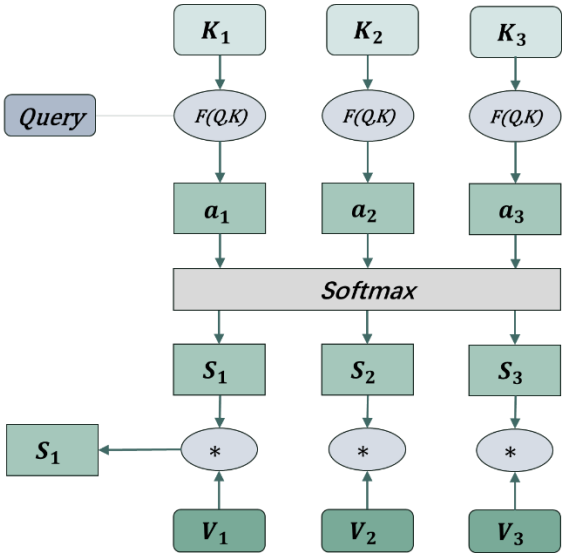

图 2-4 注意力机制

注意力的计算过程如下：

（1）计算相似度：通过点积计算查询与键的相似度，如公式（2.10）所示：

$$score(Q, K) = \frac{QK^T}{\sqrt{d_k}} \tag{2.10}$$

其中，d_k是健向量的维度，用于缩放点积，防止数值过大或过小。

（2）计算注意力权重：通过 softmax 函数将相似度转换为权重，如公式（2.11）所示：

$$\alpha = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2.11}$$

其中，权重 α 表示每个历史时间步对目标时间步的重要性。

（3）加权求和：使用权重对值进行加权求和，如公式（2.12），得到输出：

$$Attention(Q,K,V) = \alpha V \tag{2.12}$$

2.1.3.1 自注意力机制

自注意力机制（Self-Attention）是一项可捕捉序列数据内部关系的技术手段。借助该机制，模型在处理数据之际，可动态地去关注序列里不同位置间的关联，并不需要依靠外部的干预。在自注意力机制里，查询、键以及值均源自同一个输入序列，模型计算每一个时间步的查询和其他时间步的键之间的相似度，生成注意力权重，之后对所有时间步的值进行加权求和操作，以此获得每个时间步的输出表示，如公式（2.13）：

$$Self - Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.13}$$

自注意力机制处理时间序列的核心优势体现在其有全局依赖建模以及动态特征聚焦的能力，通过计算序列中任意两个时间步的关联权重，它可突破传统循环神经网络存在的局部视野限制，同步捕捉如宏观经济周期的长期趋势与如突发事件冲击的短期波动之间的复杂关联，其动态权重分配机制可以自适应地强化关键时间点的特征贡献，抑制噪声干扰。注意力机制并不依赖时序递归计算，借助矩阵并行运算可大幅提高训练效率，特别适用于高频金融数据的实时预测需求。

2.1.3.2 多头注意力机制

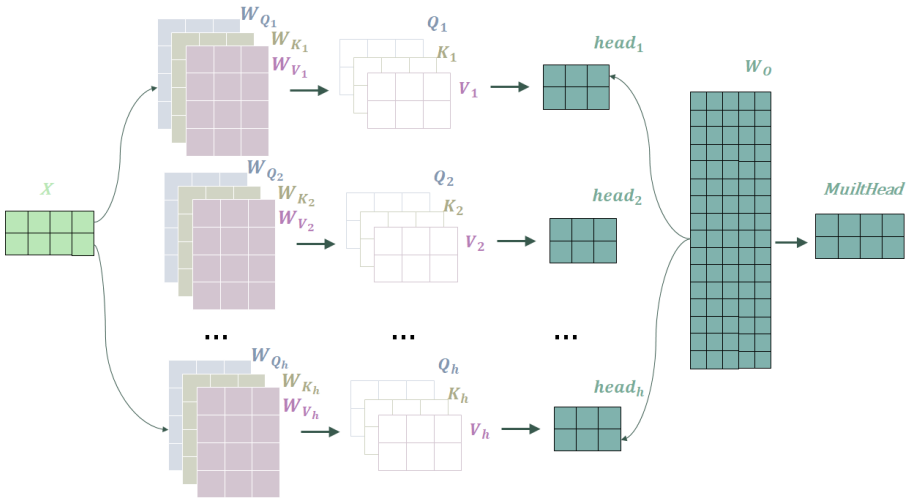

图 2-5 多头注意力机制

在多头注意力机制中，查询、键和值同样是来自同一个输入序列，但运用多个独立的线性变换来生成多个头的查询、键和值。如图 2-5，模型依靠多个头并

行计算注意力，每个头关注不同的特征子空间（如一个头关注价格，另一个头关注成交量），最后把所有头的输出拼接，得到每个时间步的表示。在时间序列预测这项任务里，这些机制可帮助模型更有效地捕捉股票数据当中的复杂依赖关系，提升预测的精准度。

每个头的查询、键和值通过独立的线性变换生成，如公式（2.14）所示：

$$Q_i = XW_{Q_i}, K_i = XW_{K_i}, V_i = XW_{V_i} \tag{2.14}$$

其中，W_{Q_i}、W_{K_i}、W_{V_i}是第i个头的可学习权重矩阵，每个头关注输入序列的不同子空间，从而捕捉更丰富的特征表示。

多头注意力机制的计算过程如下：

（1）计算每个头的注意力机制输出，如公式（2.15）所示：

$$head_i = Attention(Q_i, K_i, V_i) \tag{2.15}$$

（2）拼接所有头的输出，如公式（2.16）所示：

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W_O \tag{2.16}$$

其中，W_O是输出层的权重矩阵。

2.1.3.3 交叉注意力机制

交叉注意力机制（Cross-Attention）是一种基于注意力机制的交互式建模办法，它的核心思想是借助动态权重构建不同特征空间之间的关联映射，达成多模态数据的信息融合。和自注意力机制（Self-Attention）以及多头注意力机制（Multi-Head Attention）相比较，交叉注意力机制里的查询（Query）和键值（Key/Value）来自不一样的特征序列，以挖掘不同特征空间里潜在的协同关系或者对抗关系。

构建交叉注意力机制首先要把来自两个特征空间的查询序列和键值序列分别进行线性投影（公式（2.9）），将它们对齐至统一维度，接着缩放点积以计算注意力权重矩阵，以此量化不同序列特征间的相关性，依据权重矩阵对值向量进行加权聚合，生成融合了上下文信息的新表征。

交叉注意力机制在多模态融合任务以及跨序列建模中有着广泛应用，它的动态权重分配以及显式关系建模为复杂场景下的特征融合提供了通用且高效的解决办法。

2.2 情感分析技术

2.2.1 基于金融情感词典的情感分析方法

情感分析（Sentiment Analysis）算法用于识别文本中的情感倾向，给文本做出分类标注或者情感评分，辅助理解文本所表达的情感。基于情感词典的方法借助构建包含情感标注的词典，对每个词汇的情感色彩给予标注，并依照特定规则确定文本的情感倾向。而基于机器学习或深度学习的方法把情感分析当作分类问题，借助提取文本中代表情感倾向的特征并训练分类器来实现文本分类。

基于词典的情感分析方法是一项被广泛运用的技术，其核心在于利用预先构建好的情感词典库，对文本中的情感词进行权值统计分析，判断句子的情感极性。

基于词典的文本情感分析流程如图 2-6 所示。

图 2-6 文本情感分析流程

（1）分词：运用 PKUSEG 库对句子开展分词处理，PKUSEG 是由北京大学语言计算与机器学习研究组研发的一款高效且准确度高的中文分词工具包，它在中文自然语言处理领域表现突出，广泛应用于文本分析、情感分析、命名实体识别等任务。

PKUSEG 的核心原理是基于统计学习和深度学习技术，结合大规模语料库进行训练，它融合了基于规则和基于统计的方法，利用语言学规则和词典进行初步分词，再借助隐马尔科夫模型（HMM）或者条件随机场（CRF）等机器学习算法，优化分词结果。PKUSEG 针对不同领域，比如新闻、医学、社交媒体，提供了专门的预训练模型。这些模型依靠领域特定语料库展开训练，可更有效地识别

领域内专业术语与语言特点。用户可借助自定义词典来扩展 PKUSEG 的词汇表，以此提升特定领域的分词准确率，比如在金融领域，用户可添加“市盈率”“资产负债表”等专业术语。PKUSEG 在多个公开数据集上的分词准确率高于其他工具，在处理复杂句子与专业术语时表现颇为突出。

（2）去除停用词：停用词指在文本中频繁出现但对文本内容分析贡献不大的词汇，例如“的”“是”“在”等，去除这些词可减少数据维度，提高后续处理的效率与效果。首先要定义一个停用词列表，该列表囊括了需从文本中去除的词汇，停用词列表能依据语言和特定领域需求进行定制。接着遍历分词后的词汇列表，过滤停用词，重新组合成文本。

（3）情感词识别：首先分别获取含有正面词和负面词列表的情感词典，然后从去除停用词的词汇列表中提取出正面词（Positive Word）和负面词（Negative Word），并分别计算其数量记为PW_{num}和NW_{num}。

（4）计算情绪指数（emotion index），如公式（2.17）所示：

$$emotional_{index} = \frac{PW_{num} - NW_{num}}{PW_{num} + NW_{num}} \tag{2.17}$$

构建情感词典是情感分析流程中很关键的环节，因其直接影响分析结果的准确性。在股票领域，新闻文本往往更加精炼，大多是对客观事件的总结与报道，涉及专业术语较多，表达情感比较隐晦，情感极性不明显。投资者评论一般包含丰富情感信息，但也会出现一些比较口语化的专业性词汇，这给情感分析带来一定挑战。在构建情感词典时要保证词典中包含足够的股票领域专业术语，以便准确捕捉文本中的情感信息，还需收集投资者评论中可能出现的口语化专业性词汇，要对这些表达进行适当处理与标注。

2.2.2 基于 SnowNLP 库的情感分析方法

SnowNLP 是一个基于 Python 的中文自然语言处理（NLP）库，专注于中文文本的处理和分析。它有分词、词性标注、情感分析、文本分类、关键词提取等功能，广泛应用于中文文本的情感分析、舆情监测等领域。SnowNLP 的核心算法基于概率模型和机器学习方法，主要包括以下部分：

（1）情感分析：SnowNLP 的情感评分原理结合了情感词典加权和机器学习分类模型（默认为 SVM），依靠两阶段协同工作实现对中文文本情感倾向的量化评估。第一步先依据情感词典进行基础评分，SnowNLP 内部存有大约 20,000

个中文情感词，每个词都被标注了正向或者负向的权重，此权重是依据语料库中情感词的出现频率以及情感强度进行动态调整的，高频且情感强烈的词会拥有更高的权重。第二步运用机器学习模型展开分类，首先把文本转换为向量表示（TF-IDF 或 Word2Vec），以此捕捉上下文关系，接着借助预训练的 SVM 模型将文本分类为正面或者负面，并输出概率值，最后将分类概率线性映射到 [-1, 1] 区间，得到最终的情感得分。

（2）关键词提取：使用 TF-IDF（词频-逆文档频率）算法，评估词语在文本中的重要性，提取关键词。

（3）文本分类：将文本按预设类别（如新闻、科技、体育）分类，或按照情感（正面、负面、中性）分类。

Snownlp 简单易用、中文支持、功能丰富和可扩展性，但也存在依赖词典、性能较低、准确率有限和维护更新较慢等缺点。本文将该方法与基于情感词典的情感分析方法进行对比，探究其对股票预测的影响。

2.3 本章小结

本章主要探讨了基于深度学习模型和情感分析的股票价格预测方法中涉及的关键理论和方法，对这些理论和方法进行了原理分析和客观评价，为后续的算法实现和模型对比提供了有力的理论基础。首先介绍了时间序列预测任务常用的深度学习模型，包括卷积神经网络（CNN）、长短期记忆网络（LSTM）以及注意力（Attention）机制，这些技术方法在处理时间序列任务中有各自的优势和缺点，在本文提出的 CNN-LSTM-Attention 多层次特征融合模型中，结合了这些方法的优点。其次介绍了两种用于情感分析的方法，在后面的实验中，将对比两种情感分析方法的结果，本章的介绍为融合模型的设计以及优化提供了指导依据。

第三章 基于多层次特征融合的股票预测模型

3.1 模型设计与实现

3.1.1 模型框架

在股票预测中，传统方法如 ARIMA、GARCH 等统计模型对非线性特征的捕捉能力存在一定限度，而单一的深度学习模型往往具有局限性，如卷积神经网络 CNN 在提取序列的局部特征方面有一定的优越性，但池化操作容易丢失时间维度上的连续性；长短期记忆网络 LSTM 擅长进行序列的长期依赖建模，然而对窗口内多特征之间的关联敏感度不够，并且当参数规模过大时会致使训练效率变低。

本文提出一种基于多层次特征融合的股票预测模型 CNN-LSTM-Attention，旨在解决传统模型在股票预测中的局限性。该模型的设计目标包含三点：其一，要将 CNN 有的局部特征提取能力以及 LSTM 擅长的长期依赖建模优势相结合；其二，借助注意力机制，动态地聚焦于关键的时间步，以此提高对市场中突发信号的响应能力；其三，优化计算效率，让其可高效处理大规模交易数据，契合对历史行情的深度分析与预测需求。该模型主要由三个核心模块组成：空间特征提取、时间依赖建模、动态权重分配。

（1）空间特征提取：采用三层一维卷积网络，卷积核宽度为 5，滤波器数量分别是 32、64、128，依靠局部卷积操作，这个模块可捕捉数值数据在短期内的模式。在每层卷积之后，引入最大池化操作，来压缩序列长度，同时保留关键的波动特征。利用 ReLU 激活函数提高该模块的非线性表达能力，提高对复杂特征模式的学习。参数共享机制可有效降低计算复杂度，并且在一定程度上减轻过拟合问题，利于提高模型的泛化能力。

（2）时序依赖建模：搭建两层单向的 LSTM 网络，实现对股价长期变化规律的深度挖掘。第一层 LSTM 设置 128 个处理单元，先对数据进行初步处理，提取较为基础、局部的时序特征；第二层 LSTM 设置 256 个处理单元，增加隐状态维度，基于第一层提取的基础特征，挖掘更复杂、更高阶的特征，提升模型对复杂模式的捕捉和表征能力。逐步递增单元的方式，能保证模型学习到足够复杂的特征，又能有效避免一开始就使用过多单元导致计算资源过度消耗和训练时间过长，在模型性能和计算效率之间取得较好平衡，可降低过拟合风险和提高泛化能

力。

（3）动态权重分配：引入多头自注意力机制，借助缩放点积来计算时间步之间的关联权重。多头并行机制可以分别学习不同的特征子空间，如价格、交易量等不同特征的交互模式，突破单一特征建模的局限性，最后将多头输出拼接起来，并依靠全连接层映射为统一的表示。注意力权重矩阵可量化各个时间步对预测目标的关键性，提高模型的可解释性。

3.1.2 模型实现

本文提出的 CNN-LSTM-Attention 多层次特征融合模型结构如图 3-1 所示。

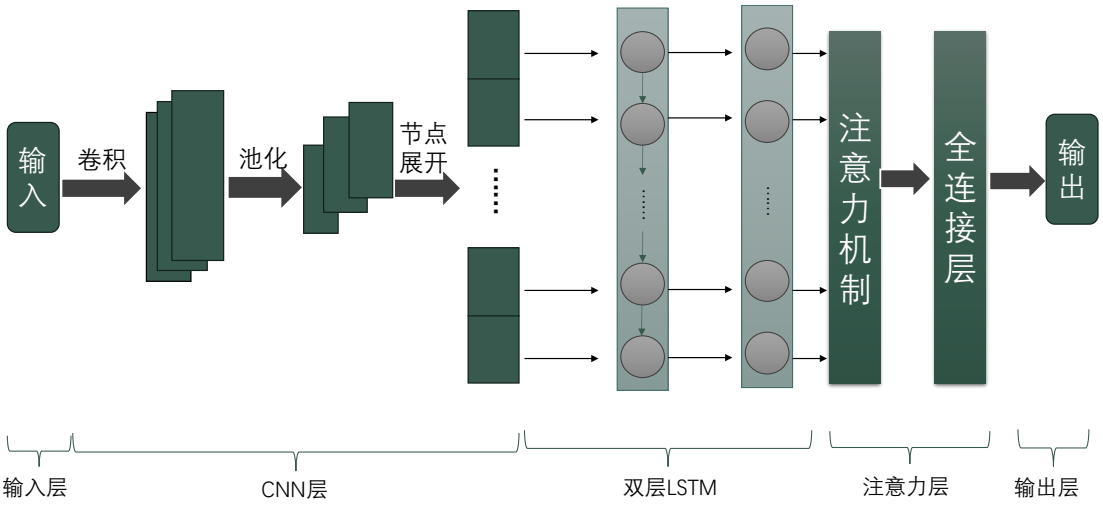

图 3-1 基于多层次特征融合的股票预测模型结构图

模型各网络层如下：

（1）一维卷积层（CNN1D）

一维卷积网络层依靠滑动卷积核捕捉局部时序窗口内的特征交互模式，自动学习数值数据的短期规律，其中卷积核可滤除噪声，提取更高层次的特征。参数共享机制降低计算复杂度，多级卷积实现多尺度特征融合。计算如公式（3.1）：

$$H^{(c)} = \sigma(W_c * X + b_c) \tag{3.1}$$

其中，$X \in \mathbb{R}^{T\times F}$为输入序列，T 为回溯天数，F 为特征个数，$W_c \in \mathbb{R}^{k\times F}$为卷积核权重，k 为核宽度，$*$为卷积操作，$\sigma$为非线性激活函数（常用 ReLU、Tanh）。

该层负责提取局部窗口内特征的时序相关性，运用多层嵌套卷积构建多尺度特征表示。

（2）长短期记忆网络（LSTM）

长短期记忆网络层通过长期依赖建模记忆历史信息，捕捉股价的长期特征，通过隐藏状态编码了截至当前时刻的全局信息。模型中叠加 2 个 LSTM 层，增强模型的复杂模式学习能力。

（3）多头注意力机制层（Muti-Head Attention）

多头注意力机制可实现动态权重分配，借助注意力机制识别关键时间步（如股价突变点），为关键时刻赋予更高权重。多头机制可并行学习不同子空间的关联（如价格与成交量之间的交叉影响），以此提升信息整合能力。该层让时序关系得以增强，直接对任意两个时间步的关系进行建模，缓解了 LSTM 的长程依赖衰减问题。

首先输入要经过线性变换，即对输入序列进行 h 次独立的线性投影（h 为注意力头数），生成查询（Query）、键（Key）、值（Value）矩阵，分别如公式（3.2）（3.3）（3.4）：

$$Q = W_Q H^{(b)} + b_Q \in \mathbb{R}^{T \times d_{model}} \tag{3.2}$$

$$K = W_K H^{(b)} + b_K \in \mathbb{R}^{T \times d_{model}} \tag{3.3}$$

$$V = W_V H^{(b)} + b_V \in \mathbb{R}^{T \times d_{model}} \tag{3.4}$$

其中，$W_Q, W_K, W_V \in \mathbb{R}^{D \times d_{model}}$为可学习的投影矩阵，$b_Q, b_K, b_V \in \mathbb{R}^{d_{model}}$，为偏置项。

接着，要缩放点积注意力，每个头独立计算注意力权重，并对键向量进行缩放，如公式（3.5）：

$$Attention_i(Q, K, V) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \in \mathbb{R}^{T \times d_k} \tag{3.5}$$

其中，$Q_i, K_i, V_i \in \mathbb{R}^{T \times d_k}$，表示第 i 个头的查询、键和值，$\sqrt{d_k}$是防止梯度消失的缩放因子。

最后，要将所有头的输出拼接，如公式（3.6）：

$$\widehat{H} = Concat(Attention_1, Attention_2, \ldots, Attention_h) W_O \in \mathbb{R}^{T \times d_{model}} \tag{3.6}$$

其中，$W_O \in \mathbb{R}^{h \cdot d_k \times d_{model}}$为拼接后的全连接权重矩阵

多头注意力机制依靠多个独立注意力窗口并行运作，捕捉数值数据中的复杂关系以及不同的特征模式，可降低预测误差，提高推理速度。在关键时间点，多

头注意力机制可自动调整权重，提高对重要信息的关注。多维度的注意力模式分析，提升了时间序列预测任务的可解释性，自动学习全局与局部的关系，使整体模型的长程依赖建模能力更为突出。

（4）全连接输出层

全连接层接收注意力层输出的张量，经过与权重矩阵的计算，将高维特征压缩成一维预测，输出要预测的收盘价，如公式（3.7）所示。该层借助L_2正则化约束权重分布，防止出现极端预测值。

$$\hat{Y} = W_y \cdot \hat{H} + b_y \tag{3.7}$$

其中，$\hat{H} \in \mathbb{R}^{T \times d_{model}}$为多头注意力输出，$W_y \in \mathbb{R}^{d_{model} \times 1}$为全连接权重。

该层完成了回归任务，输出连续型预测值，同时借助L_2正则化约束预测值分布。

本模型把 CNN 的局部特征提取能力与 LSTM 的时序建模优势进行纵向堆叠，同时引入横向注意力机制，形成空间特征提取、时序建模、动态注意力机制三个阶段的协同，实现多尺度特征融合，即空间局部尺度（CNN）、时间序列尺度（LSTM）、语义关联尺度（Attention）。该模型借助卷积核权重共享策略压缩参数量，提高训练速度，利用注意力机制动态分配特征权重，以提高模型对关键市场信号的响应能力，同时避免传统级联结构的梯度弥散问题。

3.2 实验设置

3.2.1 数据集获取与预处理

本实验所用数据集均从 Tushare 平台获取。Tushare 是由国内开发者社区打造的开源金融数据平台，为学术研究、量化投资以及金融数据分析提供免费且全面的数据支持。平台汇聚了沪深交易所、港交所、美股等市场的结构化数据，覆盖股票、基金、期货、期权等十余类金融产品，数据时间跨度从 2000 年至今，日均更新频率达 T+1 级别。Tushare 所有数据均可免费用于学术研究及非商业用途，这种开放性打破了传统金融数据平台（如 Wind、同花顺）在数据获取权限和费用上的壁垒，适合学生及个人研究者开展探索性研究。平台提供标准化的 CSV/Excel 格式数据文件，便于数据读取与模型输入。

沪深300指数（CSI-300 Index）是从上海证券交易所和深圳证券交易所主板、

创业板及科创板中选取的 300 只规模大、流动性强、代表性高的 A 股股票组成的综合性市场基准指数，可以有效反映沪深市场中上市公司的整体表现。本实验选取了数据集沪深 300 综合指数（CSI-300 Composite Index）从 2011 年 1 月 17 日至 2021 年 12 月 30 日的股票交易数据。数据集划分为训练集与测试集，分别包含 1935 个交易日和 728 个交易日的基本价格与成交量信息。本实验筛选出在 2011 年 1 月 17 日至 2021 年 12 月 30 日交易天数占比超过 98%的股票，最终构建包含 88 只股票的数据集。

数据集特征如表 3.1 与表 3.2 所示。

表 3.1 实验数据集特征（1）

date	open	close	high	low	volume
2010/1/4	0.350400411	0.305859242	0.351257814	0.338396436	0.095873498
2010/1/5	0.339396796	0.300570268	0.341540359	0.325106413	0.220539749
2010/1/6	0.332251623	0.295410234	0.332251623	0.324677693	0.163331485
2010/1/7	0.327249939	0.2921853	0.329393539	0.320104766	0.140819272
2010/1/8	0.321533771	0.291540223	0.325106413	0.319390245	0.114349029
2010/1/11	0.335824191	0.291540223	0.338396436	0.318389923	0.175498978
2010/1/12	0.32253413	0.289605198	0.323677371	0.311530514	0.234527517

表 3.2 实验数据集特征（2）

date	dopen	dclose	dhigh	dlow	dvolume	price
2010/1/4	0.3504	0.305859	0.351258	0.338396	0.095873	7.741297
2010/1/5	-0.011	-0.00529	-0.00972	-0.01329	0.124666	7.607433
2010/1/6	-0.00715	-0.00516	-0.00929	-0.00043	-0.05721	7.476833
2010/1/7	-0.005	-0.00322	-0.00286	-0.00457	-0.02251	7.39521
2010/1/8	-0.00572	-0.00065	-0.00429	-0.00071	-0.02647	7.378883
2010/1/11	0.01429	0	0.01329	-0.001	0.06115	7.378883
2010/1/12	-0.01329	-0.00194	-0.01472	-0.00686	0.059029	7.329907

本文分别将数据集大小设置为 5、25、50、75、88，探究该模型的股票预测

效果。

3.2.2 模型参数设置

输入数据借助滑动窗口生成，窗口长度为 50 日，步长为 1 日，对特征进行标准化处理后输入模型。

在进行卷积网络设置的过程中，选用了核宽度为 k=5 的设置方式，其目的在于捕捉 5 日交易周期内的特征协同情况（如量价关系、均值回归），同时采用了 3 层卷积层，以多级嵌套的形式来提取日级、周级以及月级的多尺度模式，滤波器数量分别设定为 32、64、128，以此逐步扩大特征表征空间，实现计算与表达能力的平衡。

在设置 LSTM 时，采用了双层堆叠的机构形式，选用的单元数分别为 128 和 256，通过增加单元数量以及网络深度，促使模型可逐步学习从基础特征到高级特征的抽象表示，这提升了模型容量，还提高了泛化能力。金融数据存在强自相关性，时间序列中的历史值对当前值有着明显影响，这种影响容易致使模型过拟合局部模式。为此采用 dropout 策略，随机屏蔽 30%的神经元输出，以此抑制因金融数据强自相关性而导致的过拟合现象。

在设置多头注意力机制时，将头数设定为 8，8 个头如此便可从价格、交易量、涨跌幅等多个特征子空间里，充分挖掘不同维度特征之间的交互模式。每头的维度d_k设置为 32，这样做可让模型可以更全面地捕捉数据特征，又不至于因维度过高而出现计算冗余或者过拟合的情况。

在最后的输出层，对全连接权重采用L_2正则化，这样可以将预测值约束在合理区间，防止极端预测（如股价预测超出合理范围）。

模型各网络层参数设置如表 3.3 所示。

表 3.3 模型参数设置

网络层	参数	设置值
CNN1D	核宽度 k	5
	卷积层数	3
	滤波器数量 C	32→64→128
	激活函数	ReLU
LSTM	单元数量	128

续表 3.3 模型参数设置

LSTM	Dropout 率	0.3
Multi-Head Attention	头数 h	8
	每头维度d_k	32
Output Layer	全连接权重L_2正则化	λ=0.005

本实验采用 Adam 优化器，损失函数采用MSE，初始学习率设置为$1 \times e^{-3}$，衰减率为 0.95/epoch，自适应的学习率加速收敛，适合小样本金融数据。批次大小设置为 64，采用早停机制，连续 3 个 epoch 验证机损失不降则终止训练。

3.3 评价指标

均方误差（MSE）：预测值与实际值之差的平方的平均值，如公式（3.8）。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.8}$$

其中：n 是样本数量，y_i是第 i 个样本的实际值，$\hat{y}_i$是第 i 个样本的预测值，MSE 能够量化预测误差的大小，且由于对误差进行了平方，因此更加关注较大的误差；MSE 的值越小，表示模型的预测性能越好。

平均绝对误差（MAE）：预测值与实际值之差的绝对值的平均值，如公式（3.9）所示。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.9}$$

其中：n 是样本数量，y_i是第 i 个样本的实际值，$\hat{y}_i$是第 i 个样本的预测值，MAE 同样能够量化预测误差的大小，但对异常值的敏感度较低，MAE 的值越小，表示模型的预测性能越好。

平均绝对百分比误差（MAPE）:预测误差的绝对值与实际值之比的平均值，如公式（3.10）所示。

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.10}$$

其中：n 是样本数量，y_i是第 i 个样本的实际值，$\hat{y}_i$是第 i 个样本的预测值，MAPE 能够提供误差的相对大小，对于实际值很小的情况尤其有用；MAPE 的

值越小，表示模型的预测性能越好。MAPE 为 0%表示完美模型，MAPE 大于 100%表示劣质模型。

涨跌准确率：通常用于衡量金融预测模型在预测资产价格涨跌方向上的准确性，通过与相邻元素的差值来计算涨跌方向，涨跌准确率=预测正确的次数/涨跌总次数，该指标越高则说明模型性能越好。

3.4 实验结果与分析

本文采用目前流行的深度学习框架 TensorFlow 和 Keras 进行实验。为了验证本文提出的多层次特征融合模型 CNN-LSTM-Attention 具有良好的性能，将从以下两个方面进行对比分析。一是数据集对比，从选出的 88 只股票中分别随机挑选 5，25，50，75 只股票合成的数据集，以及 88 只股票合成的数据集，在这些数据集上进行评估对比 MAE、MSE 等评价指标；二是模型对比，将本文提出的 CNN-LSTM-Attention 融合模型与常用的预测模型 CNN、RNN、LSTM、GRU 进行对比。不同数据集大小时，模型的预测性能分别如表 3.4-3.8 所示。

表 3.4 数据集大小为 5 时各模型股票预测结果

模型名称	MAE	MSE	MAPE(%)	跌涨准确率
CNN	0.02062	0.00048	30.95754	0.47619
RNN	0.02808	0.00082	66.57511	0.43260
GRU	0.02759	0.00081	22.67446	0.46948
LSTM	0.01759	0.00058	68.65416	0.40040
CNN-LSTM-Attention	0.01570	0.00043	21.90276	0.54614

表 3.5 数据集大小为 25 时各模型股票预测结果

模型名称	MAE	MSE	MAPE(%)	跌涨准确率
CNN	0.01664	0.00032	28.30207	0.49095
RNN	0.01483	0.00033	49.26248	0.48616
GRU	0.01730	0.00064	37.46823	0.46718
LSTM	0.01327	0.00014	57.62088	0.44981
CNN-LSTM-Attention	0.01206	0.00021	19.42807	0.53526

表 3.6 数据集大小为 50 时各模型股票预测结果

模型名称	MAE	MSE	MAPE(%)	跌涨准确率
CNN	0.01302	0.00017	25.58856	0.49168
RNN	0.01040	0.00055	72.64018	0.48514
GRU	0.01029	0.00062	32.95051	0.45755
LSTM	0.00948	0.00012	94.16073	0.46483
CNN-LSTM-Attention	0.00909	0.00014	22.11625	0.52967

表 3.7 数据集大小为 75 时各模型股票预测结果

模型名称	MAE	MSE	MAPE(%)	跌涨准确率
CNN	0.00512	0.00003	50.73734	0.50521
RNN	0.01477	0.00302	44.89376	0.46291
GRU	0.00597	0.00006	61.30783	0.45983
LSTM	0.00494	0.00008	59.90056	0.46682
CNN-LSTM-Attention	0.00819	0.00031	15.91019	0.53457

表 3.8 数据集大小为 88 时各模型股票预测结果

模型名称	MAE	MSE	MAPE(%)	跌涨准确率
CNN	0.00611	0.00142	15.36958	0.48592
RNN	0.00703	0.00025	77.52081	0.47454
GRU	0.00638	0.00010	62.96272	0.48209
LSTM	0.00470	0.00006	76.92535	0.46433
CNN-LSTM-Attention	0.00544	0.00006	19.34925	0.52326

表 3.4-3.8 记录了不同数据集的输入在不同股票价格预测模型上的各项评价指标。由表格可以看出，本文提出的多层次特征融合模型 CNN-LSTM-Attention 在 MAE、MASE、MAPE、涨跌准确率这些指标上综合性能最好，相对于其他几个模型误差最小，跌涨准确率最高。而且随着数据量的增大，CNN-LSTM-Attention 模型的各项指标均有提升。这有力地证实了该融合模型在股票预测上的

优势，其中的 CNN 网络可以弥补 LSTM 和 Attention 机制对空间模式的不敏感性，而 LSTM 通过建立长期依赖，缓解了 CNN 对时间动态变化的建模不足，Attention 动态分配权重，突破了传统模型均匀处理时间步的局限。由此可得，该模型可以结合各板块的优势，在股票预测中可以取得更好的效果。

3.4 本章小结

本章提出了基于多层次特征融合的股票预测模型，针对金融时序数据的非线性与高噪声特性，通过 CNN 模块进行局部特征提取，利用卷积核自适应学习不同时间尺度下的波动模式；通过 LSTM 网络建模时间序列的长短期依赖关系，门控机制保留关键信息；多头注意力机制通过权重矩阵动态量化多源信息的贡献度，提升模型对市场信号的敏感性。实验表明，多层次特征融合的模型在不同数据集大小下进行股票预测任务时优于 CNN、RNN、LSTM、GRU 等单一模型，具有较低的误差和更高的涨跌准确率，预测值与实际值更加接近。

第四章 基于市场情绪融合的股票预测模型

4.1 模型设计与实现

4.1.1 模型框架

在金融市场价格变动机制里，投资者情绪和市场数据有着紧密联系，单一数值预测模型在极端行情下解释能力不足，社交媒体情绪、新闻舆情等非结构化信号对市场参与者决策的驱动作用不能被忽视。在情绪因子与股票数值数据融合时，仍存在一些挑战，数值数据如价格、交易量等有很强的时序关联和平稳性，而情绪数据呈现出突发性和稀疏性，直接进行特征拼接容易引发模态冲突。且情绪与市场价格的因果关系存在动态滞后结构，设计模型时要捕捉跨模态滞后效应。

本文在多层次特征融合模型 CNN-LSTM-Attention 的基础上，引入了情绪特征，构建了基于市场情绪融合的股票预测模型。该模型目标包括：独立地提取数值特征与情绪特征在各自领域的特定模式；通过交叉注意力机制，对数值信息与情感信息之间的动态关联进行建模；实现数值信息与情感信息关联可视化。该模型主要由四个核心模块组成：数值特征分支、情感特征分支、特征融合、特征拼接与回归。

（1）数值特征分支：采用 CNN 与 LSTM 的级联结构，借助多尺度卷积与递归网络共同挖掘数据的局部特征以及长期依赖。三层递增通道（32→64→128）的一维卷积网络可实现从局部波动到全局趋势的数据特征提取，把原始数据转变为高阶时序表征，参数共享机制让模型对特征出现的时序位置保持稳健性。双层 LSTM 网络能解析长程关联，逐层递增的单元数（128→256）逐步扩充记忆容量，捕捉数值序列的长期依赖关系，有效缓解梯度消失问题。

（2）情绪特征分支：和数值数据不同，从新闻及评论中获取的情绪数据的有高噪声、稀疏性与突发性等特性，该模块采用“轻量化 CNN-LSTM”架构。两层一维卷积网络采用窄核卷积，核宽度分别设为 5、3，通道数逐层增多(16 → 32)。窄核卷积在保留情绪数据突变信号的同时能有效过滤高频噪声，聚焦于局部时间步的情绪波动，可精准捕捉短周期内情绪倾向的剧烈变化。逐层递增的通道数设计，能实现情绪特征的分层抽象。轻量化的结构设计极大降低了模型参数量和计算复杂度，能有效防止过拟合现象，同时加快模型的训练与推理过程。

（3）特征融合模块：数值特征有平滑趋势性和物理意义明确的量纲，而情

绪特征呈现稀疏性和语义模糊性，传统物理拼接方式仅从数据结构层面进行简单的整合，却忽略了二者在信息熵维度的巨大差异，会让两类特征信息相互干扰。本模型引入交叉注意力机制来进行特征融合，建立数值与情绪特征的动态关联映射，以可解释的注意力权重量化情绪因子与数值变动之间的关联。

情绪数据一般包含大量噪声，作为查询（Query）可能致使模型过度关注噪声相关的数值波动，降低鲁棒性，当某段时期情绪与数值的关联性较弱时，情绪作为查询会使预测结果严重偏离实际。在常态市场中，数值指标如价格、交易量等具有更强的趋势延续性，作为查询可提高模型在平稳期的预测精度，并且数值数据一般都经过清洗以及标准化的处理，可让注意力机制更加专注于那些与当前价格走势逻辑相契合的情绪信号，避免受到情绪噪声的干扰。以数值数据的长期序列作为查询内容，模型可对各指标的周期性波动进行建模。本模型采用数值数据当作查询，把情绪数据作为键值。

运用交叉注意力机制来进行特征融合，可以提高模型的动态关联建模能力，也就是可自适应地去计算数值与情绪特征的逐日关联权重，在市场波动剧烈的时期，可强化对突发情绪信号的捕捉，而在市场平稳的时期，则可以降低噪声干扰，达成数值特征与情绪特征的动态适配。交叉注意力机制的设计还可有效地规避情绪噪声干扰，以数值数据作为查询，可引导注意力机制优先去关注与数值数据变化逻辑一致的情绪信息，过滤掉情绪噪声，让模型更加聚焦于有预测价值的情绪特征。注意力权重的显式计算，将情绪因子对数值变动的影响程度量化成为可视化指标，为模型推理过程提供直观的语义解释，提升股票价格预测的可解释性。

（4）特征拼接与回归：截取数值分支与注意力模块输出的末端进行拼接，通过全连接层网络输出价格预测值。注意力模块在某时间步的输出表示数值分支在该时间步的查询结果，依靠加权聚合情绪分支所有时间步的情绪特征所得到的上下文向量，将数值分支与其查询结果的最后一个时间步进行拼接，模型可捕捉情绪历史对当前数值状态的影响，动态地选择与当前数值状态最相关的情绪历史信息。

在回归层部分采用了两个线性层，和单层线性回归相比，双层结构可建模特征间的非线性组合，同时分阶段降维可以让梯度在两次线性变化中逐步得到调整，避免因为输入与输出维度差异过大而导致的梯度波动。借助 ReLU 激活函数缓解

梯度消失问题，同时稀疏化特征表示，提高模型的泛化性能。

4.1.2 模型实现

本文提出的基于市场情绪融合的股票预测模型结构如图 4-1 所示。

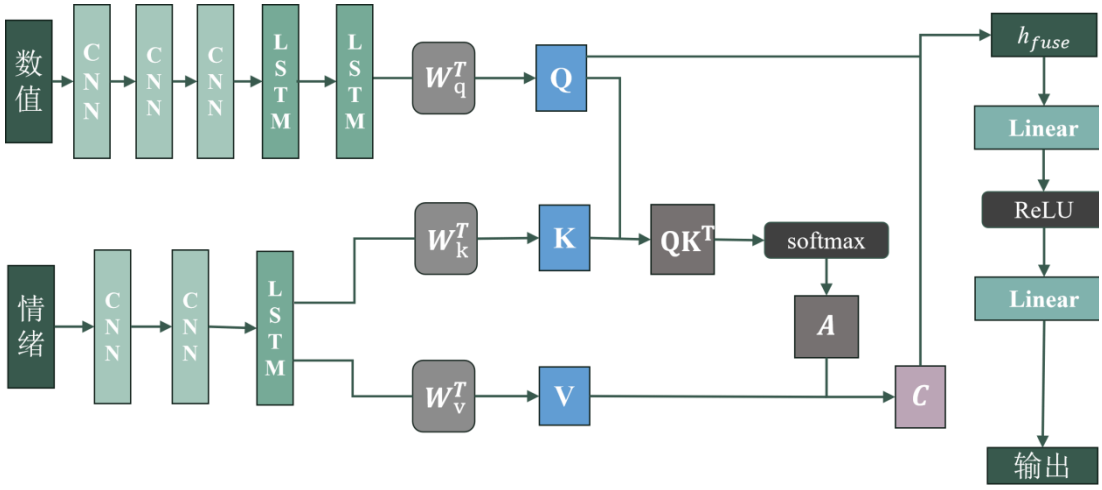

图 4-1 基于市场情绪融合的股票预测模型结构图

模型各网络层如下：

（1）一维卷积层（CNN1D）：一维卷积层依靠卷积核在时间轴上滑动，关注局部时间段内的数据特征，可自动识别出数据中的短期规律。因为参数共享，同一个卷积核在整个时间序列上重复使用，可以提升计算效率。堆叠多个卷积层使网络可融合不同时间尺度的特征，这种层次化的处理使得模型可更全面地理解数据中的复杂模式。

在数值分支模块中，采用了三层堆叠的一维卷积网络，卷积核宽度设置为 5，通道数逐层扩大（32→64→128），这种设计可捕获不同时间尺度的局部特征，从短期波动到中期形态再到长期周期模式。

在情绪分支模块中，采用了轻量化设计，包含两层一维卷积网络（16→32），卷积核宽度设置为 5、3，这种设计可有效过滤情绪数据中的噪声，同时保留关键的情绪突变信号。

（2）长短期记忆网络（LSTM）：LSTM 依靠独特的门控机制，可有效处理时间序列数据里的长期依赖关系，减轻传统循环神经网络在训练时出现的梯度消失问题，借助细胞状态在序列里传递信息，与门控机制共同发挥作用，达成对信息的选择性保留与更新，精准捕捉数据在长时间跨度下的变化趋势。

在数值分支模块当中，运用两层级联的 LSTM 结构（128→256），第一层初步捕捉数值序列中的短期依赖模式以及基础时序特征，第二层挖掘更深的长期依赖关系，依靠单元数递增的设计让模型构建从微观到宏观形成的多层次时序表征。

在情绪分支模块里，采用一层 64 维的 LSTM 层，接收第二层一维卷积网络输出的特征，对情绪信号进行有效整合，对情绪传播过程中的时序变化进行建模，捕捉情绪在不同时间点的演变规律以及相互关联。

（3）交叉注意力机制：采用交叉注意力机制来进行特征融合，以挖掘数值特征与情绪特征之间的动态关联，把数值特征当作查询，情绪特征作当作键值，依靠计算查询与键之间的相似度得分，对值进行加权求和，实现特征之间的交互融合。

首先依靠线性投影对齐维度，使查询与键值可交互，如公式（4.1）-（4.3）：

$$Q = W_Q H^{num} + b_Q, W_Q \in \mathbb{R}^{128\times64} \tag{4.1}$$

$$K = W_K H^{sent} + b_K, W_K \in \mathbb{R}^{64\times64} \tag{4.2}$$

$$V = W_V H^{sent} + b_V, W_V \in \mathbb{R}^{64\times64} \tag{4.3}$$

然后计算数值与情绪的跨模态关联，如公式（4.5）所示：

$$A = Softmax\left(\frac{QK^T}{\sqrt{64}}\right) \in \mathbb{R}^{B\times50\times50} \tag{4.5}$$

其中A的每个元素a_{ij}表示第i个时间步的数值特征与第j个时间步的情绪特征的相关性。

接下来通过加权聚合情绪特征，如公式（4.6）所示：

$$C = AV \in \mathbb{R}^{B\times50\times64} \tag{4.6}$$

其中 C 编码了情绪特征对数值特征的修正信号。

该机制可按照数值特征的变化情况，动态挑选出与之关联紧密的情绪特征，对情绪因子对于数值变动的影响程度给予量化，构建起两种特征之间有效的映射关系。

（4）特征拼接与回归层：该层承担着把融合后的特征加以整合，输出最终预测结果的任务。它截取数值分支 LSTM 输出的最后一个时间步特征以及交叉注意力模块输出的最后一个时间步特征，如公式（4.7）将这两个特征进行拼接，得到结合了数值信息与融合信息的特征向量。之后借助两层全连接网络把特征映

射到预测目标维度，如公式（4.8）和（4.9）以此完成回归预测，双层线性结构分阶段降低维度，逐步调整梯度，这样可避免因为输入输出维度差异过大而致使的梯度波动，提升模型的预测精度以及泛化性能。

$$h_{fuse} = \left[H^{num}[:,-1,:]; C[:,-1,:]\right] \in \mathbb{R}^{B\times 192} \tag{4.7}$$

$$z = ReLU\left(h_{fuse}W_{l1} + b_{l1}\right), W_{l1} \in \mathbb{R}^{192\times B} \tag{4.8}$$

$$\hat{y} = zW_{l2} + b_{l2}, W_{l2} \in \mathbb{R}^{64\times 1} \tag{4.9}$$

其中W_{l1}, W_{l2}为权重矩阵。

本模型运用双分支 CNN-LSTM 网络，分别对数值数据与情绪数据展开处理，提取其中的空间特征以及时序规律，以此来保留各自领域的特征独立性，随后引入交叉注意力机制，借助该机制模型得以自动识别情绪和技术指标的关联强度随着时间窗口所呈现的动态变化，达成股票数值特征与情绪特征的协同建模。

4.2 实验设置

4.2.1 数据集获取与预处理

4.2.1.1 数据收集

东方财富网是国内领先的金融信息服务平台，覆盖超过 90%的 A 股上市公司信息，其股吧用户活跃度很高，评论量达千万级/日，具备天然的流量入口优势。图 4-2 是东方财富网股吧论坛的帖子，有标注帖子的时间戳和发表来源。

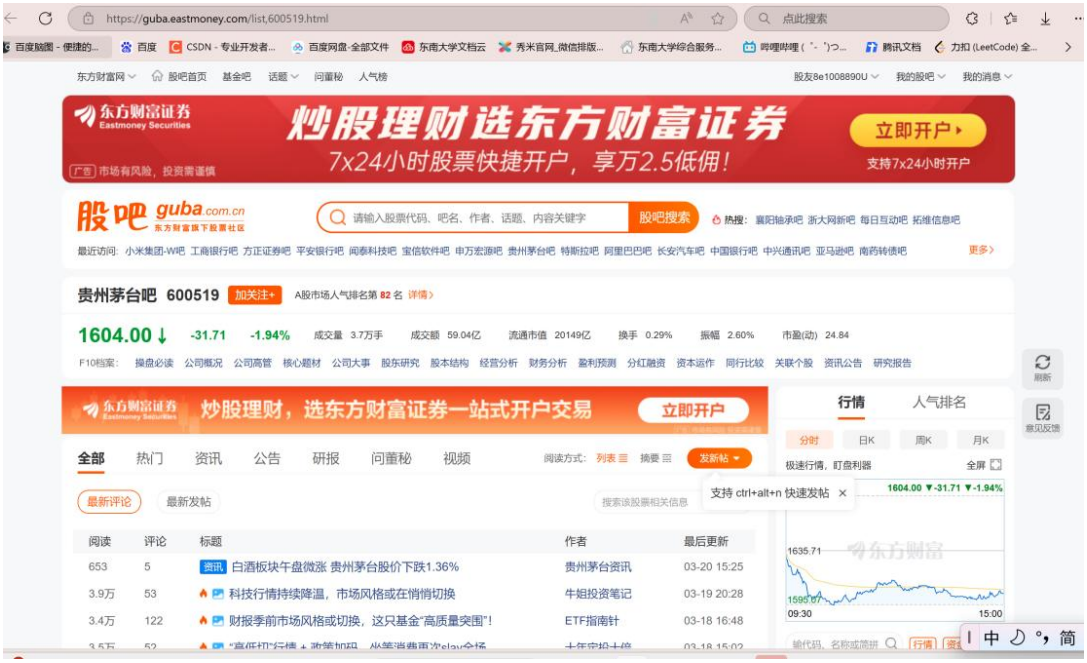

图 4-2 东方财富网股吧

该平台既聚合了证券公司研报、上市公司公告等权威信息，又包含了个体投资者的评论数据，形成“专业机构+散户投资者”的多层级信息生态。其中机构动向类新闻的情感极性变化（如“增持”“减持”关键词）可直接映射到机构资金流向，散户讨论贴也呈现明显的情绪传染特征，有助于提取情绪信息，进行情感分析，为股票预测提供更多维的信息。该平台的帖子更新速度快、用户活跃度高，有助于及时获取投资者对股市变动的情感态度和观点倾向；除此之外，该平台的帖子带有时间戳，有利于情感信息与股票数值数据的时间对齐，从而更精准地研究投资者情感与股票价格变化之间的联系，提高股票价格预测的准确性。

本实验主要从个股层面出发，使用 python 的 selenium 库进行动态爬虫，爬取了贵州茅台（600519）、工商银行（601398）、小米集团-W（hk01810）三只股票的股吧帖子，获取各股票新闻和评论的标题、网页链接和时间，时间跨度如表 4.1 所示。

表 4.1 各股票数据起止时间

股票代码	起始时间	终止时间	交易日天数
600519	2018-04-19	2025-01-17	1640
601398	2019-03-08	2025-01-02	1415
hk01810	2018-07-09	2025-01-16	1608

4.2.1.2 数据处理

本文使用 python 的 selenium 库进行动态爬虫，获取新闻和评论的标题、网页链接和时间，然后再根据爬到的网页链接爬取新闻的内容，通过 sumy 库生成摘要，将新闻和评论进行时间对齐后保存在 csv 文件中，过程如图 4-3。

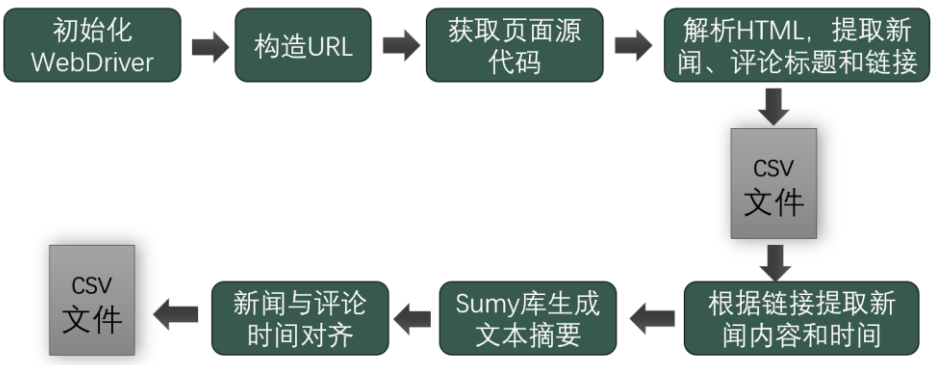

图 4-3 股票新闻和评论爬取

原始新闻和评论生成摘要的过程只是缩小文本的体量，但内容仍然比较杂乱，因此我们需要对这些文本数据进行预处理操作，便于后续的实验，过程如图 4-4。

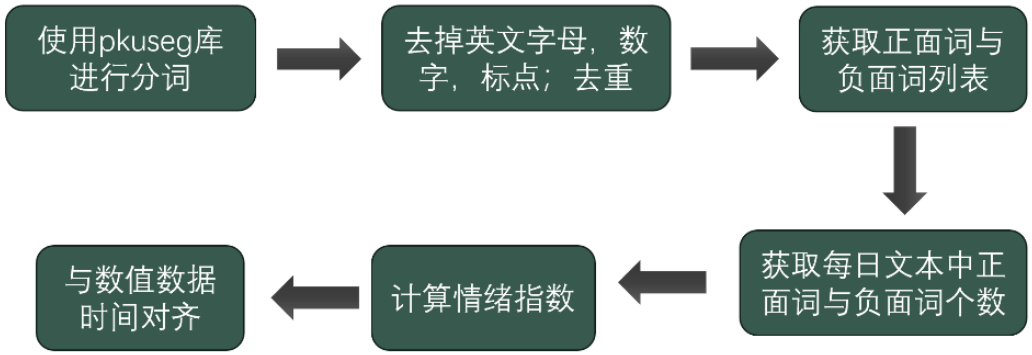

图 4-4 文本处理

分词的作用是把连续的中文文本划分成有意义的词语单元，以此来处理金融领域的术语识别问题，如表 4.2。运用 pkuseg 分词工具对新闻以及评论数据开展分词工作，加载其内置的金融领域词典，此工具可达成多模式切分策略，还可以自动识别组织机构名（如“贵州茅台”）、时间表达式及复合词（如“护盘行为”）。

表 4.2 分词示例

分词前	茅台多地冰淇淋门店闭店 白酒年轻化之路何去何从 茅台全年出口营收首次突破 50 亿元 白酒，怎么喝才健康 巨龙抬头欲腾飞！万辰集团走出十分完美的巨龙腾飞形态，说明庄家有坚如磐石的信心和高 贵州茅台跌成高股息股，消费板块今年怎么看？
分词后	茅台|多|地|冰淇淋门店闭店|白酒|年轻化|之|路|何去何从|茅台|全年|出口|营| 收|首次|突破|50 亿元|白酒|，|怎么|喝|才|健康|巨龙|抬头|欲|腾飞|！|万辰集 团|走出|十分|完美|的|巨龙|腾飞|形态|，|说明|庄家|有|坚如磐石|的|信心|和|高|贵州|茅台|跌成|高|股息|股|，|消费|板块|今年|怎么|看|？

分词得到的文本数据会存在噪声，比如评论中的表情、符号、换行符等，它们会影响情感分析的效果，降低效率，因此，我们需要对这些文本进行清洗，消除文本中的噪声干扰，提升后续分析效率，如表 4.3。

表 4.3 数据清洗示例

清洗前	茅台|多|地|冰淇淋门店闭店|白酒|年轻化|之|路|何去何从|茅台|全年|出口|营|收|首次|突破|50 亿元|白酒|，|怎么|喝|才|健康|巨龙|抬头|欲|腾飞|！|万辰集 团|走出|十分|完美|的|巨龙|腾飞|形态|，|说明|庄家|有|坚如磐石|的|信心|和|高|贵州|茅台|跌成|高|股息|股|，|消费|板块|今年|怎么|看|?
清洗后	茅台|多|地|冰淇|淋门|店|闭|店|白酒|年轻化|之|路|何去何从|茅台|全年|出口|营|收|首次|突破|亿元|白酒|怎么|喝|才|健康|巨龙|抬头|欲|腾飞|万辰集团|走 出|十分|完美|的|巨龙|腾飞|形态|说明|庄家|有|坚如磐石|的|信心|和|高贵州| 茅台|跌成|高|股息|股|消费|板块|今年|怎么|看

去除停用词的目的在于过滤无实际含义的高频虚词，提升特征表达能力。停用词包括基础停用词、金融专用停用词和动态停用词，如表 4.4。

表 4.4 停用词示例

基础停用词	“的”“了”“在”等通用虚词
金融专用停用词	“机构”“财报”“个人”等非情感载体词汇
动态停用词	随时间、事件或语境变化而临时失去信息价值的词汇，如统计语料中高频无意义词（如“真的”“绝对”），按特定出现频率自动剔除。

4.2.1.3 情绪指数计算

我们采用计算情绪指数这一方法来进行情感分析。首先要建立情感分析的正面词和负面词列表，我们将姜富伟等人提出的中文金融情感词典列表[27]作为我们的基准词汇表。在中文金融情感词典的基础上，我们还搜集了金融领域口语化的专业表达和术语变体，分别加入到情感词典的正负面词汇列表中，提高情绪指数的准确性。然后从预处理好的文本中提取所有词语，与预设的金融情感词典进行匹配，统计每日文本中正面词与负面词的数量PW_{num}与NW_{num}，并根据公式（4.10）计算情绪指数。

$$emotional_{index} = \frac{PW_{num} - NW_{num}}{PW_{num} + NW_{num}} \tag{4.10}$$

在将情绪指数与股票数值数据进行时间上的对齐之后，采用格兰杰检验获取情绪相对于数值数据的最佳滞后，变换之后再输入模型进行训练。

4.2.2 模型参数设置

在数值特征提取模块当中，首先运用三层一维卷积网络来开展多尺度特征提取，每一层的卷积核宽度都设定为 5，其目的在于捕捉短交易周期之内的特征，多级嵌套的卷积层可提取不同时间长短的多尺度模式，滤波器数量依次设置为32、64、128，以此逐步扩大特征表征空间，平衡计算效率与表达能力。激活函数选用 ReLU，以此提高非线性建模能力，双层 LSTM 的隐藏单元数分别为 128、256，借助逐步扩大隐藏单元维度的方式，系统性地捕捉时序数据的层次化特征表达，可保留数据的微观波动，还强化了对宏观趋势的建模能力。

在情绪特征提取模块当中，采用的是双层卷积结构，第一层的卷积核宽度被设定为 5，其作用是提取情绪趋势里相对宏观的波动情况，而第二层的核宽度则设置成 3，它更加侧重于局部细节方面，滤波器的数量分别被设置成 16 以及 32，以轻量化设计来适配情绪数据所有的稀疏性特点。后续的 LSTM 层隐藏单元数设定为 64，以此来捕捉更为复杂的时序特征，提高时序建模的能力。

在交叉注意力模块，把数值分支 LSTM 输出的 256 维数据当作查询，将情绪分支输出的 64 维数据作为键值，这个模块可让数值特征动态地选择相关的情绪上下文，减轻情绪与数值特征之间的错位问题。回归预测采用了两级全连接结构，第一层借助 ReLU 激活函数达成非线性交互建模，第二层直接输出价格预测。输出层运用L_2正则化方法把预测值限制在合理的区间范围内。

在最后的输出层，针对全连接权重采用L_2正则化，如此可将预测值约束在合理区间，避免出现极端预测的情况。模型参数设置如表 4.5。

表 4.5 模型参数设置

模块	网络层	参数	设置值
数值特征提取	CNN1D	核宽度 k	5
		卷积层数	3
		滤波器数量 C	32→64→128
		激活函数	ReLU
	LSTM	单元数量	128→256
		Dropout 率	0.3
情绪特征提取	CNN1D	核宽度 k	5→3

续表 4.5 模型参数设置

情绪特征提取	CNN1D	卷积层数	2
		滤波器数量 C	16→32
		激活函数	ReLU
	LSTM	单元数量	64
		Dropout 率	0.3
数值特征与情绪特征融合	交叉注意力	Query 维度	256
		Key/Value 维度	64
回归预测	全连接层	全连接维度	320→64→1
		L_2正则化系数	λ=0.001

本实验选用 Adam 优化器，损失函数采用MSE，初始学习率设置为$1 \times e^{-3}$，衰减率为 0.95/epoch，自适应的学习率加速收敛，适合小样本金融数据。批次大小设置为 32，采用早停机制，连续 3 个 epoch 验证机损失不降则终止训练。

4.3 评价指标

本章节评价指标采用均方误差（MSE）、平均绝对误差（MAE）、平均绝对百分比误差（MAPE）以及涨跌准确率，计算公式见章节 3.3。

4.4 实验结果与分析

交叉注意力机制可以生成注意力权重热力图来可视化情绪特征与数值特征之间的协同关系，图 4-5 的热力图呈现出了在股票 601398 验证集中情绪特征与数值特征所存在的动态关联模式，其中横轴代表着情绪序列时间步，纵轴则表示数值序列时间步，而颜色的强度体现出了注意力权重值的大小情况。

以图 4-5 为例进行分析，颜色趋势大致呈对角线分布但有所偏移，说明情绪与数值特征存在明显的协同作用，印证了前期数据处理时通过格兰杰检验法选择最佳情绪滞后的合理性，也表明交叉注意力机制在进一步调整情绪滞后方面具有重要作用。左下角权重值更大表明模型更加关注早期历史事件步的特征间关联；在颜色强度低的带状区域内（情绪 30-38，数值 13-20），权重值较小表明这段时间内情绪与数值特征的协同作用有限。整体来看，此模型更依赖长期历史特征协同，而对近期特征间关联关注有限。该图为模型在验证集上预测时情绪特征与数

值特征关联的整体情况，将数据集每个批次数据预测时单独绘制热力图则可得到更小时间尺度内情绪特征与数值特征之间的相互关联，这对于定位重大市场事件、分析股票价格变动因子、预测股票走势具有重要作用。

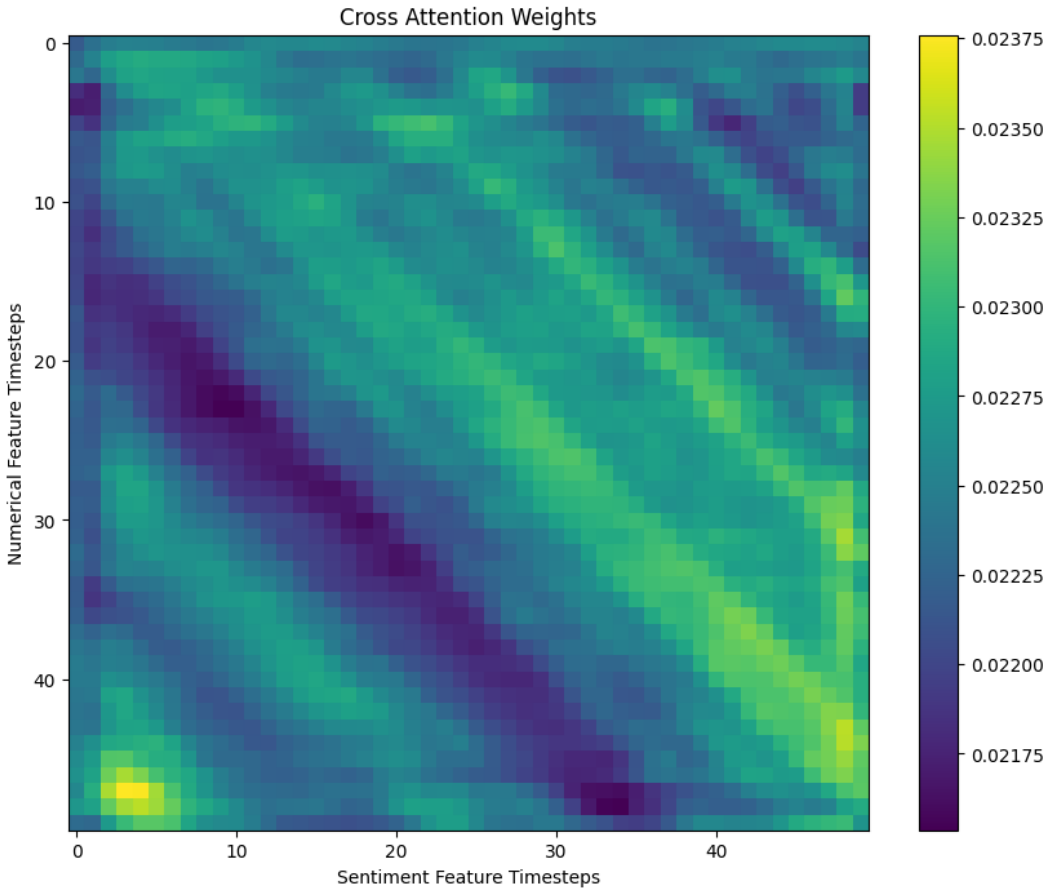

图 4-5 股票 601398 验证集数值与情绪特征交叉注意力权重热力图

为了验证情绪因子对于股票预测有正面影响，该部分的实验将从三个方面进行对比分析。

一是情绪融合模块对股票预测的影响，通过对比融合模型是否采用了情绪融合模块来进行分析，本文选用了三只不同的股票，从东方财富网股吧获取它们的新闻与评论数据，并将其转换为情绪因子，选取不同的股票是为了避免不同类型股票数据的差异性对实验结果的影响，帮助我们更全面的分析情绪因子在股票预测中的作用；二是情绪信息对不同模型预测性能的影响，这部分将本文提出的基于市场情绪融合的股票预测模型与基础深度学习模型 CNN、RNN、LSTM、GRU 模型的预测结果进行对比分析，同时本文还在一些 transformer 系列模型上进行了预测，包括 Transformer、iTransformer、Reformer、Informer、Flashformer、

Flowformer，旨在探究不同模型对于情绪因子的加入呈现在股票预测上的影响；三是情绪信息获取方法对股票预测的影响，一方面使用计算情绪指数的方法，另一方面使用 SnowNLP 内置的情绪评分方法。

4.4.1 情绪融合模块对股票预测的影响

表 4.5 情绪融合模块对股票预测的影响

股票代码	是否采用情绪融合模块	MAE	MSE	MAPE(%)	涨跌准确率
600519	否	0.00992	0.00015	2.20629	0.52681
	是	0.00716	0.00012	1.22961	0.54351
601398	否	0.01595	0.00025	2.97203	0.43750
	是	0.01145	0.00023	2.20474	0.44118
hk01810	否	0.01263	0.00036	4.58796	0.52019
	是	0.00895	0.00032	2.88490	0.55055

表 4.5 记录了三只股票分别在未采用情绪融合模块的预测模型和基于市场情绪融合的预测模型上进行股票预测时的各项评价指标。由表格可以看出，情绪信息对于股票价格的预测有一定的正面影响。基于市场情绪融合的预测模型在测试集上的误差明显降低，而且涨跌准确率有显著提升，表明预测效果得到了提升。该实验结果证实，股票价格走向受多种因素影响，除价格等数值信息外，关注多个方面的因素如新闻情绪、投资者情感等，有利于捕捉不同影响因素之间的相互作用，促进模型对于市场复杂特征的理解，进而获得更好的预测效果。

4.4.2 情绪信息对不同模型预测性能的影响

表 4.6 股票 600519 数值+情感数据在各模型上的表现

股票代码	模型	MAE	MSE	MAPE(%)	涨跌准确率
600519	CNN	0.01208	0.00026	2.15278	0.50000
	RNN	0.00846	0.00012	1.40592	0.51899
	LSTM	0.00846	0.00012	1.40592	0.51899
	GRU	0.00771	0.00013	1.38522	0.52633
	CNN-LSTM-Attention	0.00716	0.00012	1.22961	0.54351

表 4.7 股票 601398 数值+情感数据在各模型上的表现

股票代码	模型	MAE	MSE	MAPE(%)	涨跌准确率
601398	CNN	0.02021	0.00071	5.62302	0.43667
	RNN	0.01608	0.00042	3.45432	0.40519
	LSTM	0.01279	0.00030	3.38659	0.42222
	GRU	0.01256	0.00029	3.33186	0.43481
	CNN-LSTM-Attention	0.01145	0.00023	2.20474	0.44118

表 4.8 股票 hk01810 数值+情感数据在各模型上的表现

股票代码	模型	MAE	MSE	MAPE(%)	涨跌准确率
hk01810	CNN	0.01257	0.00028	3.71185	0.52452
	RNN	0.01297	0.00029	5.38896	0.48742
	LSTM	0.01015	0.00038	5.18716	0.46806
	GRU	0.01255	0.00029	5.73042	0.48387
	CNN-LSTM-Attention	0.00895	0.00032	2.88490	0.55055

由表 4.6、4.7、4.8 可得，对该三只股票进行价格预测，本文所提出的基于市场情绪融合的预测模型在以上模型中误差最低，涨跌准确率最高，在股票价格预测任务中有较好的表现，说明融合模型能够结合各网络层的优势，提高预测性能。

表 4.9 股票 600519 两种数据在-former 系列模型上的表现

模型	数据集	MAE	MSE	MAPE(%)	涨跌准确率
Transformer	数值数据	0.05243	0.00473	35.69672	0.49235
	数值数据+情感信息	0.04253	0.00340	33.53918	0.49541
iTransformer	数值数据	0.04197	0.00368	41.35617	0.51682
	数值数据+情感信息	0.04782	0.00468	36.05464	0.49235
Reformer	数值数据	0.09273	0.01827	81.55397	0.46789
	数值数据+情感信息	0.08988	0.01209	39.84735	0.51682
Informer	数值数据	0.21042	0.06612	45.49601	0.49541
	数值数据+情感信息	0.12711	0.02652	42.74597	0.53211
Flashformer	数值数据	0.05326	0.00471	48.49660	0.48624

续表 4.9 股票 600519 两种数据在-former 系列模型上的表现

Flashformer	数值数据+情感信息	0.04546	0.00397	36.41864	0.49847
Flowformer	数值数据	0.05132	0.00472	50.41866	0.45566
	数值数据+情感信息	0.04981	0.00443	40.81914	0.51070

表 4.10 股票 601398 两种数据在-former 系列模型上的表现

模型	数据集	MAE	MSE	MAPE(%)	涨跌准确率
Transformer	数值数据	0.22314	0.13560	36.77115	0.45390
	数值数据+情感信息	0.21520	0.12036	19.86397	0.47844
iTransformer	数值数据	0.12026	0.02981	13.16281	0.39362
	数值数据+情感信息	0.11972	0.02980	12.48591	0.37589
Reformer	数值数据	0.18984	0.08729	21.38524	0.42553
	数值数据+情感信息	0.16420	0.04961	17.91124	0.43908
Informer	数值数据	0.30242	0.20454	37.25505	0.43617
	数值数据+情感信息	0.28477	0.18601	32.37983	0.45390
Flashformer	数值数据	0.28117	0.20332	22.20228	0.45035
	数值数据+情感信息	0.24838	0.15282	19.95342	0.46745
Flowformer	数值数据	0.24996	0.15594	25.96557	0.44326
	数值数据+情感信息	0.19343	0.09936	15.63372	0.46454

表 4.11 股票 hk01810 两种数据在-former 系列模型上的表现

模型	数据集	MAE	MSE	MAPE(%)	涨跌准确率
Transformer	数值数据	0.10304	0.02675	49.04487	0.48125
	数值数据+情感信息	0.07719	0.01320	45.40362	0.49500
iTransformer	数值数据	0.06757	0.00894	48.30862	0.50313
	数值数据+情感信息	0.07610	0.01035	48.87521	0.47500
Reformer	数值数据	0.09760	0.01869	64.70810	0.49375
	数值数据+情感信息	0.07468	0.01083	50.51041	0.50000
Informer	数值数据	0.27055	0.11250	60.95858	0.50938

续表 4.11 股票 hk01810 两种数据在-former 系列模型上的表现

Informer	数值数据+情感信息	0.25563	0.08090	39.98225	0.51875
Flashformer	数值数据	0.12025	0.02963	68.40622	0.47500
	数值数据+情感信息	0.07953	0.01340	41.40954	0.49813
Flowformer	数值数据	0.10988	0.02345	61.36096	0.47188
	数值数据+情感信息	0.09003	0.02029	54.53747	0.48438

表 4.12 三只股票在-former 系列模型上加入情绪信息后的平均变化

股票代码	MAE 平均降低	MSE 平均降低	MAPE 平均降低	涨跌准确率平均增长
600519	10.87%	19.42%	20.00%	6.30%
601398	9.61%	20.77%	21.73%	2.79%
hk01810	15.57%	28.88%	18.87%	1.32%

表 4.9、4.10、4.11 记录了三只股票输入数值、数值与情感两种数据在-former 系列模型上的表现。由表 4.12 可得，不同的股票数据集在加入情绪信息后性能提升各有不同，对贵州茅台（600519）数据集，加入情绪信息后涨跌准确率升高最为明显，平均增长 6.303%；对工商银行（601398）数据集，MAPE 误差降低最多，平均降低 21.729；对小米集团（hk01810），情绪信息的加入使得 MAE 和 MSE 降低显著，分别平均降低 15%和 28%。虽然对于不同股票的数据集情绪信息带来的性能增长有所不同，但整体上都使模型的股票预测性能有所提升。

表 4.13 -former 系列模型上加入情绪信息后在三只股票上预测的平均变化

模型名称	MAE 平均降低	MSE 平均降低	MAPE 平均降低	涨跌准确率平均增长
Transformer	15.84%	30.00%	19.82%	2.96%
iTransformer	-8.70%	-14.30%	5.60%	-1.79%
Reformer	13.35%	39.68%	29.78%	4.97%
Informer	16.98%	32.35%	17.85%	4.44%
Flashformer	20.06%	31.78%	24.83%	3.73%
Flowformer	14.54%	18.63%	23.32%	6.51%

由表 4.13 可得，情绪信息针对不同模型预测性能所产生的影响呈现出十分

突出的差异状况，并且这种差异和模型架构本身所有的特性有着紧密的关联。在误差指标这一方面，Reformer 以及 Informer 呈现出最为突出的优化成效，它们的 MSE 相对降低的幅度分别为 39.68%以及 32.35%，这是由于二者的注意力机制设计所导致的：Reformer 借助局部敏感哈希（LSH）可高效地将情绪特征和局部股价模式进行对齐，而 Informer 的 ProbSparse 注意力则可筛选关键的情绪事件，抑制长尾噪声带来的干扰。Flashformer 实现 MAE 最大降幅（20.06%），这证实了 IO 感知算法对于多模态数据实时融合有着增益的作用。在方向性预测中，Flowformer 以 6.51%的涨跌准确率提升领先，它的流守恒机制可动态地对情绪权重加以调节，更加契合市场突变场景下实时预测的需求。然而，iTransformer 的注意力机制与情绪时序特征之间存在兼容性方面的问题，这使得 MAE/MSE 分别恶化了 8.70%/14.30%，需要对跨模态融合结构进行改进。情绪信息对误差指标的改善效果（MAE/MSE 平均相对降低 15.8%/27.3%）普遍优于涨跌方向预测（平均提升 3.7%），这说明情绪因子更适用于股价波动幅度的建模，而涨跌方向性的判断则需要结合技术指标以及微观结构分析。

4.4.3 情绪信息获取方法对股票预测的影响

表 4.14 股票 600519 采用不同情绪分析方法在各模型上的表现

模型名称	情绪计算方式	MAE	MSE	MAPE(%)	跌涨准确率
CNN	SnowNLP	0.02381	0.00087	4.01483	0.48101
	情绪指数	0.01208	0.00026	2.15278	0.52331
	无	0.01247	0.00032	2.46915	0.51779
RNN	SnowNLP	0.01002	0.00017	1.73741	0.51899
	情绪指数	0.00846	0.00012	1.40592	0.52709
	无	0.00871	0.00013	1.42062	0.51899
GRU	SnowNLP	0.00907	0.00015	1.63921	0.54430
	情绪指数	0.00771	0.00013	1.38522	0.55633
	无	0.01050	0.00017	1.82061	0.55063
LSTM	SnowNLP	0.00862	0.00013	1.50496	0.50633
	情绪指数	0.00849	0.00013	1.48351	0.51300
	无	0.00900	0.00014	1.56488	0.50633

续表 4.14 股票 600519 采用不同情绪分析方法在各模型上的表现

	SnowNLP	0.00802	0.00089	1.26381	0.52681
CNN-LSTM-Attention	情绪指数	0.00716	0.00012	1.22961	0.54351
	无	0.00992	0.00016	2.20629	0.52681

表 4.15 股票 601398 采用不同情绪分析方法在各模型上的表现

模型名称	情绪计算方式	MAE	MSE	MAPE(%)	跌涨准确率
	SnowNLP	0.02045	0.00014	5.78467	0.41889
CNN	情绪指数	0.02021	0.00013	5.62302	0.43667
	无	0.03593	0.00028	7.48263	0.40161
	SnowNLP	0.01332	0.00033	3.57474	0.40741
RNN	情绪指数	0.01108	0.00032	3.45432	0.40919
	无	0.01418	0.00036	3.82940	0.39259
	SnowNLP	0.01415	0.00036	3.70976	0.43704
GRU	情绪指数	0.01256	0.00029	3.33186	0.45481
	无	0.01888	0.00053	5.22187	0.41481
	SnowNLP	0.01329	0.00033	3.61162	0.42963
LSTM	情绪指数	0.01279	0.00030	3.38659	0.45022
	无	0.02205	0.00067	6.44094	0.42222
	SnowNLP	0.01997	0.00033	6.52428	0.43750
CNN-LSTM-Attention	情绪指数	0.01145	0.00023	2.20474	0.44118
	无	0.01595	0.00025	2.97203	0.43750

表 4.16 股票 hk01810 采用不同情绪分析方法在各模型上的表现

模型名称	情绪计算方式	MAE	MSE	MAPE(%)	跌涨准确率
	SnowNLP	0.01973	0.00064	5.40063	0.52258
CNN	情绪指数	0.01257	0.00028	3.71185	0.52452
	无	0.01692	0.00048	6.26264	0.47177
RNN	SnowNLP	0.01381	0.00033	5.78626	0.47097

续表 4.16 股票 hk01810 采用不同情绪分析方法在各模型上的表现

RNN	情绪指数	0.01297	0.00029	5.38896	0.48742
	无	0.01641	0.00044	6.37208	0.48387
GRU	SnowNLP	0.01378	0.00034	6.35997	0.45806
	情绪指数	0.01255	0.00029	5.73042	0.48387
	无	0.01279	0.00030	5.89212	0.45806
LSTM	SnowNLP	0.01164	0.00026	5.04951	0.45806
	情绪指数	0.01015	0.00024	5.01716	0.46806
	无	0.01571	0.00041	7.52237	0.46452
CNN-LSTM-Attention	SnowNLP	0.01191	0.00048	5.91983	0.51768
	情绪指数	0.00895	0.00032	2.88490	0.55055
	无	0.01263	0.00036	4.58796	0.52019

表 4.14、4.15、4.16 展示了分别采用情感指数计算和调用 SnowNLP 内置函数获取情绪信息，并与数值信息一同输入到 CNN-LSTM-Attention 等模型中进行股票价格预测。由表格可以看出，采用公式计算获得的情感指数进行股票预测误差更低，且涨跌预测准确率优势显著。

表 4.17 三只股票采用不同情绪分析方法在模型上的性能平均变化

股票代码	情感分析方法	MAE 平均降低	MSE 平均降低	MAPE 平均降低	涨跌准确率 平均增长
600519	SnowNLP	-13.83%	-131.28%	-5.68%	-1.65%
	情绪指数	13.19%	15.49%	17.45%	1.63%
601398	SnowNLP	17.75%	22.13%	-3.46%	3.04%
	情绪指数	33.85%	34.69%	28.82%	6.01%
hk01810	SnowNLP	4.62%	-3.00%	3.77%	1.25%
	情绪指数	22.62%	26.56%	25.87%	4.83%
平均	SnowNLP	2.85%	-37.38%	-1.79%	0.88%
	情绪指数	23.22%	25.58%	24.04%	4.16%

表 4.17 记录了三只股票采用不同情绪分析方法在模型上的性能平均变化，进一步分析以上的实验结果：

首先，将新闻情绪和投资者情绪相融合可提升股票预测的准确程度，之如此是由于新闻情绪在某种程度上体现了市场预期，借助投资者关注度、新闻正负面关键词等，情绪因子可对市场针对突发事件比如政策变化、行业动态等的即时反应进行量化，而这弥补了传统数值数据存在滞后性的不足。

其次，不同股票对情绪敏感度也存在显著差异，贵州茅台（600519）属于消费股，此类高端白酒具有品牌效应，短期内的情绪波动对其实际价值影响有限；工商银行（601398）属于金融股，对情绪敏感度较高，融合市场情绪信息能够使其各模型的预测性能得到明显提升，其中 MAE 和 MSE 分别平均降低 33.85%和 34.69%，MAPE 平均降低 28.82%，涨跌准确率平均提升 6.01%，这与银行业受宏观经济政策（新闻）和散户情绪（降息预期）双重影响的特征相符；小米集团（hk01810）属于科技股，此类股票走向需结合产业动态（新闻）和机构投资者仓位变化（情绪）进行研判。

采用情绪指数作为情绪信息在提升模型的预测性能方面优于调用 snownlp 获得的情绪评分，原因在于金融投资者评论中存在大量行业特有的隐喻表达和术语变体，诸如“割韭菜”（散户被套）等的市场黑话，、“利好出尽”（正面消息触顶转跌）一类的情感极性反转词。除此之外，投资者还会用一些特殊的句法结构传递隐含情绪，比如双重否定句式。SnowNLP 作为通用情感分析工具，其训练语料库缺乏金融领域特异性，对于以上口语化或隐晦的专业表达，会出现很多无法做出情绪量化或误判的现象，影响情感评分的效果，基于金融情感词典的情感指数的计算充分考虑了这些情况。在中文金融情感词典的基础上，搜集了金融领域口语化的专业表达和术语变体，分别加入到情感词典的正负面词汇列表中，以提高情绪指数的准确性。

4.5 本章小结

本章提出一种基于市场情绪融合的股票预测模型，通过分别建立数值特征提取与情绪特征提取两个分支，分别采用 CNN-LSTM 网络提取特征信息，然后将数值分支输出作为查询（Query），将情绪分支输出作为键值（Key/Value），输入交叉注意力机制中，最后将数值分支的输出与交叉注意力模块的输出末端进行

拼接，通过两层线性网络输出最终的价格预测值。从实验结果可以得出，融合市场情绪信息可以提升深度学习模型股票预测的性能，与传统深度学习模型相比，CNN-LSTM-Attention 多层次特征融合模型在融合了情绪信息的股票预测任务中表现最好，相比于调用 SnowNLP 获得的情感评分，计算情感指数获得的情绪信息对于股票预测的提升效果更好。不同的股票数据集有不同的特点，对于新闻情绪和投资者情绪有不同的反应周期和响应程度，因而在不同股票数据集上的预测效果的提升也有区别。

第五章 总结与展望

5.1 总结

本文针对股票预测中存在的数据特征提取不全面以及模型解释性差等问题，提出了基于多层次特征融合的股票预测模型，在该模型中引入了多头自注意力机制，运用并行计算实现特征空间解耦，强化关键步特征，以此更有效地捕捉长期依赖关系；提出了基于市场信息融合的股票预测模型，分别提取数值特征与情绪特征，运用交叉注意力机制进行融合，达成市场情绪与数值信息的协同，提升预测的稳定性与准确性。本文的研究结论如下：

1. CNN-LSTM-Attention 多层次特征融合模型协同各个网络架构的互补特性，有效克服了传统单一模型在复杂金融时序分析中的固有缺陷。该模型将多头注意力模块嵌入特征融合层，利用多头并行计算机制实现跨时间步的动态权重分配。在与基础深度学习模型的对比实验中，此模型进行收盘价格预测时误差最小，跌涨准确率最高。通过融合卷积网络的局部特征提取能力、循环神经网络的长期依赖建模优势以及注意力机制的非线性关联分析特性，该模型提升了对股票预测的准确性。

2. 基于市场情绪融合的股票预测模型通过数值特征与情绪特征的分别提取，保持了两类特征的独立性，通过交叉注意力机制获取了两类特征之间的关联，发挥其协同作用进行股票价格预测。该模型相比单独使用数值数据进行预测的模型预测精度有所提高；与调用 SnowNLP 获得的情感评分相比，本文所用情绪指数对于股票预测准确性的提高更为显著。

3. 本文所构建的情绪指数和股票本身的特点存在较强相关性，不同类型的股票受突发事件（如政策影响）和舆论情绪影响的程度不同，消费板块股票相对更易受新闻情绪以及投资者情感的影响。互联网作为当代社会传播信息的关键途径，汇聚了数量众多的新闻与评论数据，这些数据以文本形式蕴含着丰富的情感信息，给研究新闻和投资者情绪提供了全新视角。相较于传统的数值数据，文本信息能更真实且全面地呈现股票相关的市场状况，体现投资者的情感变化。

本文的主要贡献在于：第一，提出 CNN-LSTM-Attention 多层次特征融合模型，多层次提取股票时间序列特征，缓解传统时序模型的长程依赖衰减问题，通过注意力机制聚焦关键交易日，提高了预测准确率；第二，为更好地利用情绪数

据服务股票预测，在 CNN-LSTM-Attention 多层次特征融合模型中融合市场情绪信息，先分别对数值数据与情绪数据进行特征提取，再通过交叉注意力协同预测，进一步提高了预测精度。

5.2 展望

本文所提出的基于多层次特征融合的股票预测模型相比传统时序预测模型（CNN/RNN/LSTM/GRU）显现出了性能方面的优势。这样的性能提升主要源于于空间特征提取、时序依赖建模以及动态权重分配这三个模块共同发挥作用，达成了“空间-时间-语义关联”多层次的特征融合。在这个模型的基础之上运用交叉注意力融合市场情绪信息，可提升模型性能。然而本研究在以下一些方面仍存在改进空间：

1. 本文数据采集存在一定的局限性。对于新闻和评论的采集只是局限于东方财富网股吧，并未将新浪财经、同花顺、雪球等投资社区的内容纳入进来；权威媒体维度缺失，未整合彭博社、路透社等专业媒体的深度报道以及机构研报图谱数据。未来的研究可以构建多源异构数据采集体系，设计基于知识图谱的跨平台信息融合框架，来实现政策文本、机构观点与散户情绪的三维数据协同分析。

2. 本文对股票新闻和评论进行情绪指数的计算较为简约，同时投资者的行业黑话不断更新，而情感词典难以覆盖所有该类专业用语，在计算情绪指数时仍会有误差存在，难以完全捕捉投资者情感的多样性和市场的复杂性。因此未来工作可以考虑对新闻和评论文本设置一个更加精细化的情绪指数计算方式和标准，从而提高模型的准确性。

3.本文在时间选取上相对粗糙，选取以交易日为单位，而在实际市场中，股市的波动状态、投资者的情感变化、市场的反应每时每刻都在变化。未来的研究可以选取更加精细化的时间段比如以小时或分钟为单位，对股票数值数据和情绪信息对齐与融合，研究新闻动态和投资者情绪对于股票预测的影响。

参考文献

[1] 武广林. 投资者情绪对股票收益率的影响研究[J]. 运筹与模糊学, 2024, 14(2): 352-359.

[2] BANERJEE D. Forecasting of Indian stock market using time-series ARIMA model[C]. 2014 2nd International Conference on Business and Information Management (ICBIM), 2014: 131-135.

[3] MAQSOOD A, SAFDAR S, SHAFI R, et al. Modeling Stock Market Volatility Using GARCH Models: A Case Study of Nairobi Securities Exchange (NSE)[J]. Open Journal of Statistics, 2017, 7(2): 369-381.

[4] LIN Y L, GUO H X, HU J L. An SVM-based approach for stock market trend prediction[C]. The 2013 International Joint Conference on Neural Networks (IJCNN), 2013: 1-7.

[5] REN Z, YIN J, YU Y C, et al. Stock price prediction based on optimized random forest model[C]. 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), 2022: 777-783.

[6] SOMKUNWAR R, RAO J, VARVANTE N. Stock Value Prediction Accuracy Enhancement Using CNN and Multiple Linear Regression for NIFTY[C]. 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2024: 1-7.

[7] BATHLA G, RANI R, AGGARWAL H. Stocks of year 2020: prediction of high variations in stock prices using LSTM[J]. Multimedia Tools and Applications, 2023, 82(7): 9727–9743.

[8] GUPTA U, BHATTACHARJEE V, BISHNU P S. StockNet—GRU based stock index prediction[J]. Expert Systems with Applications, 2022, 207(C): 117986.

[9] CHOUDHURY C, ALI K, CHANDA S, et al. Stock Price Prediction Using Recurrent Neural Networks and Genetic Algorithm[C]. 2024 IEEE Silchar Subsection Conference (SILCON 2024), 2024: 1-6.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[C]. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[11] WANG C J, CHEN Y Y, ZHANG S Q, et al. Stock market index prediction using deep Transformer model[J]. Expert Systems with Applications, 2022, 208: 118128.

[12] YULISTIANI R, KURNIADI F I. Stock Price Prediction With the Informer Model[C]. 2024 International Conference on Information Management and Technology (ICIMTech), 2024: 49-53.

[13] KITAEV N, KAISER L, LEVSKAYA A. Reformer: The Efficient Transformer[C]. 2020 International Conference on Learning Representations, 2020:12764-12775.

[14] DAO T, FU D, ERMON S, et al. FlashAttention: Fast and memory-efficient exact attention with IO-awareness[C]. In Advances in Neural Information Processing Systems (NeurIPS2022), 2022: 16344-16359.

[15] WU H X, WU J L, XU J H, et al. Flowformer: Linearizing Transformers with Conservation Flows[C]. International Conference on Machine Learning, 2022: 24226-24242.

[16] LIU Y, HU T G, ZHANG H R, et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting[C]. 2024 International Conference on Learning Representations, 2024:18933-18957.

[17] IRSYAD A, PRAFANTO A, FIRDAUS M B, et al. Forecasting Stocks Prices with GRU and Attention Mechanism[C]. 2024 International Conference on Electrical Engineering and Informatics (ICELTICs), 2024: 114-119.

[18] LEE J Y, YOO S J. Stock Price Prediction Using Transformer and Time2Vec[C]. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2025: 0687-0689.

[19] DONG Z H, FAN X Y, PENG Z Y. FNSPID: A Comprehensive Financial News Dataset in Time Series[C]. KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024: 4918-4927.

[20] HUANG C J, CHEN Z Y, SOLIMAN W M. Stock Price Prediction with FinBERT and RNN[C]. In Proceedings of the 7th International Conference on Algorithms, Computing and Systems (ICACS '23). 2024:74–79.

[21] WU S T, LIU Y L, ZOU Z R, et al. S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis[J]. Connection Science, 2022, 34(1): 44–62.

[22] S S B, FATIMA N S, ABBAS M. Enhancing Stock Market Prediction with an LSTM-Attention Model Integrating Macroeconomic and Sentiment Features[C]. 2024 International Conference

on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2024: 1-6.

[23] 张少军, 苏长利. 基于情绪词典和 BERT-BiLSTM 的股指预测研究[J]. 计算机工程与应用, 2025, 61(04): 358-367.

[24] 崔婷, 黄斐然. 基于情感分析大模型的股票预测：结合 GRU 和 ALBERT 的预测模型[J]. 东岳论丛, 2024, 45(02): 113-123.

[25] 徐洪峰. 基于深度学习的多源异构股票数据分析方法研究[D]. 福建省: 厦门大学信息学院, 2022.

[26] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[27] 姜富伟, 孟令超, 唐国豪. 媒体文本情绪与股票回归预测[J]. 经济学(季刊), 2021, 21(04): 1323-1344.

致　　谢

行文至此，落笔为终，写到论文最后章节，四年的本科生涯也即将结束。仿佛昨天初入校园，今日便弹指一瞬，到达本阶段的终点，目光所及之处皆是回忆，心怀万般不舍，亦是满分感激。

首先我要特别感谢吕建华老师在这次毕业论文中对我耐心而专业的指导，为我的论文选题、实验以及修改提出了很多宝贵的意见。吕老师严谨的治学态度与包容的指导风格，不仅帮助我顺利完成论文，更传授我探索问题的思维素养，这份收获将使我受益终身。

感谢学院的各位授课老师，为我系统性地搭建起专业领域的知识框架，从专业知识讲授到专题实践，从学术汇报到小组合作，让我得以在各方面有所成长和收获。 感谢历任辅导员老师，时刻关心我的学习与生活状态，鼓励我积极参与活动，为我解答学业生活以及未来规划中的困惑，让我能够更清晰和坚定地向目标前进。感谢各位教务老师，老师们的高效工作提供了便捷的学业辅助，尤其是在保研过程中，各种证明的快速办理使我有充足的时间提交简历和报名。

深深感恩我的家人，在学业规划和院校专业选择各方面始终尊重我的选择，给予包容与鼓励，为我提供了一个温和的家庭氛围，使我无后顾之忧地追逐自己的目标。感谢我的小伙伴们，倾听我的烦恼，驱散我的焦虑，给予我陪伴和鼓励，每次畅聊都能使我恢复能量，重新整装出发。

最后感谢勇敢的我自己，在四年的本科生涯中不断的尝试和探索，即使迷茫过，碰壁过，但在这个过程中慢慢认识了自己，找到自己的学习和生活节奏，在学业生涯上勤奋努力，在业余时间加强锻炼，同时发展自己的爱好。未来希望仍能保持对个人的清晰认知和对目标的不懈追求。

关关难过关关过，前路漫漫亦灿灿。

心求至善

SOUTHEAST UNIVERSITY