分类号：F832.51
学　号：2020410021

山东工商学院

硕士学位论文

基于深度学习的股票价格预测研究

Research on Stock Price Prediction Based on Deep Learning

研究生姓名　董玲
指导教师　原达教授

学科门类　工学
一级学科学位授权点名称　计算机科学与技术
论文提交日期　2023 年 6 月

摘要

股票研究人员和投资者一直致力于对股票市场进行可靠预测，进而获得投资收益最大化。股票波动受到多方面因素的影响，例如股票的历史价格数据、社交媒体舆论、投资者情绪等。股票价格和股票文本融合是一种有效的股票预测方法，但是仍然存在历史价格数据的时间依赖性差、股票文本可用性低以及融合特征有效性不足等问题。现有的股票数据中的噪声数据、低质量数据以及不完整的异常数据，导致学习到的股票特征不准确，使得模型预测性能低下。此外，现有的模型大多是通过改变股票预测的网络结构提高股票数据集的的可用性，对股票数据的不确定性因素缺少深入的研究。因此，本文从提高股票输入的可用性，充分选取股票相关特征以及选择高效预测模型入手，提升股票预测模型的性能。重点研究内容如下：

(1) 考虑到股票市场影响因素复杂多样以及股票相关数据可用性不高的问题，本文提出了一个基于异构数据和多层注意力机制的股票价格预测模型 CredibleNet，核心创新在于对股票历史价格的依赖关系进行捕获，使用单词层、句子层的两层注意力机制对股票中的低质量文本进行处理。将股票文本与股票价格两种异源数据进行融合，使用时间级注意力机制提取股票融合特征中的有价值信息，获得高效的融合数据信息。结果表明，注意力机制能够有效的解决股票数据中质量低、可信度低、信息可用性低等问题。此外，本文使用高斯混合模型对神经网络的输出进行建模，进而对股票预测任务进行不确定性量化，分析股票预测的不确定性。

(2) 针对股票数据中的异常数据造成的预测效果不理想、预测模型精度较低这一问题，本文使用极端决策树模型对股票的历史价格交易数据进行特征选择，然后将不确定性量化的思想引入到股票预测模型中，进而提升模型的准确性。在此基础上，本文提出了基于特征选择和不确定性量化的股票价格预测模型 LogNet，使用高斯混合模型对神经网络的输出的的 logit 进行建模，选取具有较低不确定性的股票数据进行股票预测。在选取预测模型时，本文选择使用 SDENet 模型对股票未来趋势进行预测，该模型能够从低不确定性的股票特征中提取信息，进而提升模型预测性能。对比实验结果表明，LogNet 相比于其他股票预测神经网络模型性能有显著的提升。

(3) 为了更加充分的利用股票的相关特征，本文提出了一个基于情感分析和 Transformer 的股价预测模型 SenTransNet。该模型首先对从推特获得的非结构化数据构建情感分数，然后将情感得分与股票的历史价格指标进行融合。在股票预测模型的选取

上，本文利用 Transformer 模型，实现对股票融合数据的有效分析，进而提升股票价格的预测性能。为了有效的评估 SenTransNet 模型的有效性和稳定性，本文选取了四个不同行业中四支不同的股票个股作为数据集，结果表明所提出的 SenTransNet 股票预测模型对于股票预测任务具有较好的预测性能和鲁棒性。

关键词：注意力机制；特征融合；不确定性；特征选择；情感分析

Abstract

Stock researchers and investors have been working hard to make reliable forecasts on the stock market and maximize investment returns. Stock volatility is affected by many factors, such as historical stock price data, social media public opinion, investor sentiment, etc. Stock price and stock text fusion is an acceptable stock forecasting method. However, problems remain, such as poor time dependence on historical price data, low stock text availability, and insufficient fused features' effectiveness. Noisy, low-quality, and incomplete abnormal data in the existing stock data lead to inaccurate learned stock features and poor model prediction performance. In addition, most current models improve the availability of stock data sets by changing the network structure of stock forecasting and lack in-depth research on the uncertainty factors of stock data. Therefore, this paper starts with improving the availability of stock input, entirely selecting stock-related features, and selecting an efficient forecasting model to improve the performance of the stock forecasting model. The critical research contents are as follows:

(1) Considering the complex and diverse influencing factors of the stock market and the low availability of stock-related data, this paper proposes a stock price prediction model CredibleNet based on heterogeneous data and multi-layer attention mechanisms. The core innovation lies in the historical price of stock dependency relationship captured, and the low-quality text in the stock is processed using the two-layer attention mechanisms of the word layer and the sentence layer. Then, the two heterogeneous data of stock text and stock price are fused, and the time-level attention mechanism is used to extract valuable information in the stock fusion features to obtain efficient fusion data information. The results show that attention mechanisms can effectively solve the problems of low quality, low credibility, and low information availability in stock data. In addition, this paper uses a Gaussian mixture model to model the output of the neural network and then quantifies the uncertainty of the stock forecast task and analyzes the uncertainty of the stock forecast. Experimental results thoroughly verify that the proposed model performs well on the tweet and stock price datasets.

(2) Aiming at the problem of unsatisfactory prediction effect and low prediction model

accuracy caused by abnormal data in stock data, this paper uses a highly randomized tree model to select features of historical stock price transaction data. Then it quantifies uncertainty. The idea is introduced into the stock forecasting model to improve the accuracy of the model. On this basis, this paper proposes a stock price prediction network LogNet based on feature selection and uncertainty quantification, uses a Gaussian mixture model to model the logit of the neural network's output, and selects stock data with lower uncertainty for stock prediction. When selecting a prediction model, this paper uses the SDENet model to predict the future trend of stocks. This model enables us to extract information from low-uncertainty stock characteristics, thereby improving the model's prediction performance. The results of comparative experiments show that LogNet has significantly improved performance compared to other stock forecasting neural network models.

(3) Investor sentiment has an essential impact on stock price fluctuations, but most existing research rarely considers the integration of sentiment values and other heterogeneous data. To solve this problem, this paper proposes a stock price prediction model SenTransNet based on sentiment analysis and Transformer. The model first constructs sentiment scores on unstructured data obtained from Twitter and then fuses the sentiment scores with historical price indicators for stocks. In selecting the stock prediction model, this paper uses the Transformer model to realize the practical analysis of the stock fusion data, thereby improving the stock price prediction performance. To effectively evaluate the effectiveness and stability of the SenTransNet model, this paper selects four different stock stocks in four other industries as data sets. The results show that the proposed SenTransNet stock prediction model has better prediction performance and robustness for stock prediction tasks.

Key Words: attention mechanism; feature fusion; uncertainty; feature selection; sentiment analysis

目 录

第 1 章　绪论

本章介绍研究背景及意义，论述股票预测领域的国内外研究现状以及存在的问题，并给出本文的研究内容和创新点。

1.1　研究背景和意义

相比于债券和储蓄存款等投资品种，股票具有更高的收益潜力，因此吸引了大量投资者的青睐。而且股票市场的信息公开透明，投资者可以通过财务报告、行业分析等方式了解市场走势和公司状况，这为投资者做出投资决策提供了依据。如何选择合适的股票预测模型实现投资收益最大化一直作为热点问题吸引着众多研究者和投资人员。随着全球经济的快速发展，股票市场规模逐渐增大，使得股票市场变的更为复杂。股票市场作为一个高度复杂的混沌系统，其趋势不仅受到历史股票价格的影响，金融环境、新闻报道、国家政策以及社会舆论等诸多因素也与股票市场的变化密切相关。股票市场的高波动性和高不确定性，使得预测股票成为一项非常艰巨的任务。为了帮助股票投资人员提供更多有效的股票投资信息，使其能够获得利润最大化，越来越多的研究者和投资者都加入到股票预测的队伍当中。提高股票预测的准确性以及可靠性，帮助投资者获取利润，规避风险，在一定程度上能够维护社会经济稳定。

然而，股票市场的高波动性和非平稳性，使股票的精准预测依旧是一个巨大的挑战。从研究数据类型来看，传统的股票预测方法主要基于时间序列模型对股票的历史价格数据进行研究[1]。人们认为所有与股票相关的信息都体现在股票价格上，因此只使用股票价格进行分析就可以预测股票市场行情。然而，仅仅使用历史价格进行预测是不准确的。股票波动受到来自技术层面、消息层面等多方面因素的影响[2]。数据源的丰富性有利于更加全面地了解股票市场，并做出比以前更为准确的预测。董理等人[3] 发现社交媒体中的评论信息对股票指数波动具有一定的影响，证明了股票评论与股票市场之间存在一定的关系。随着自然语言处理 (Natural Language Processing, NLP) 技术的发展，人们开始将社交媒体以及相关文本加入到股票预测的影响因素中。然而，与股市相关的非结构化数据质量都比较低，尤其是在线内容中有很大一部分是可信度低的谣言，这会对股票预测模型的有效性造成很大的影响。对一家公司而言，分析其综合报告要比分析对该公司的评论更有可能产生准确的预测。因此，使用非结构化数据对股票趋势进行预测的有效性与所使用内容的质量密切相关。若文本数据的可用性

不高，预测结果也会存在很大的偏差。

股票预测常见的分析方法有基础面分析、统计方法、传统机器学习和深度学习的方法。基础面分析是对影响股票价格变动的敏感因素进行分析，研究股票价格变动的一般规律[4]，但是一般预测的时间跨度比较长，对短期股票预测而言可用性不高。而统计方法是利用数理统计技术对股票指数的走势进行分析，此类方法可以在一定程度上拟合时间序列的波动变化，但是对于高度非线性、不平稳的股票数据存在一定的不适用性[5, 6]。随着机器学习的飞速发展，研究者们发现使用机器学习方法进行股票预测可以在一定程度上弥补上述方法的不足，机器学习模型利用算法不断对模型进行优化来提高股票预测的性能。但是传统机器学习的结构较为简单，模型预测结果不稳定。随着新兴技术的发展，深度学习以其记忆性、参数共享和图灵完备性等优点，引起了股票研究者们的广泛关注[7]，大量的研究表明，相比于传统的股票预测方法，利用深度学习方法进行股票预测在提高准确性等方面取得了不错的进展。

股票作为一种具有高收益、高透明度的投资产品，吸引了大量投资者的青睐，为了追求利润，股票预测成为了投资人员研究的焦点。有效的股票预测模型能够为投资者提供有关未来股市趋势的信息，降低投资风险、提高投资收益。股票市场中的任一因素的微小变化都可能会导致股票市场整体发生巨大波动，仅凭简单的分析难以准确推测股票未来趋势。

1.2 国内外研究现状

1.2.1 股票预测方法研究概述

早期的股票预测方法可以根据信息依赖的类型分为两种: 技术分析和基本面分析。技术分析主要通过分析具有时间序列性质的技术指标间的内在规律，对股票市场价格进行预测。常见的技术指标如移动平均线、随机指标以及移动平均收敛散度等[8, 9]。技术分析方法中指标类数据比较容易进行模型分析，是目前研究中常采用的数据来源[10, 11]，这些指标可以发现用于未来预测的交易模式。自回归模型[12](Autoregressive Model, AR) 是典型的技术分析方法，该模型主要针对线性和平稳时间序列进行建模。然而，股票数据常常受政策和新闻的影响而波动，股票价格的非线性和非平稳性限制了 AR 模型的适用性。为了解决这个问题，之后的研究者们尝试使用非线性学习方法来捕捉市场趋势背后的复杂模式。Nayak 等人[13] 使用由支持向量机 (Support Vector

Machines, SVM) 和 K 近邻 (K-Nearest Neighbor, KNN) 系统组成的综合模型预测用于预测股票市场的未来损益。为了进一步模拟时间序列中的长期依赖性，循环神经网络 (Recurrent Neural Network, RNN)，尤其是长短期记忆 (Long Short Term Memory, LSTM) 网络，常被用于股票预测研究[14–16]。

技术分析方法存在的缺陷是无法揭示价格数据以外的影响股票市场波动的因素。然而研究表明，仅使用价格因素或其他指标数据进行股票预测提升性能是有限的[17]。基本面分析方法则是从外部市场寻求信息，找到影响公司或行业的潜在因素作为预测因素。通过分析与股票相关的经济政策、金融环境、媒体新闻报道等因素，推断股票的现实价格高于还是低于其内在价值，从而判断股价的涨跌趋势。随着互联网技术的不断进步和普及，人们越来越多地使用社交媒体平台获取和共享信息，进而产生了大量在线内容。这些具有多样性在线信息的为研究股票市场提供了新的研究方向。新闻报道以及社交论坛上与股市相关的消息成为进行基本面分析的重要来源。为了更好地预测市场趋势，已经有很多股票研究人员尝试挖掘文本数据进而分析股票未来趋势。Nassirtoussi 等人[18] 基于突发的财经新闻文本提出了一种具有语义和情感的多层降维算法，用于预测外汇市场的方向走势。Wang 等人[19] 利用文本回归任务来预测股票价格的波动性。Hagenau 等人[20] 从财经新闻文本信息中提取大规模的表达特征来表示非结构化文本数据，并利用特征选择算法来优化股票预测进而提高股价预测精度。如果只是基于上述方法很难进一步挖掘股票数据的内部隐含信息，在预测股票变化趋势方面很难有所突破。随着股票趋势研究的深入以及与深度学习技术在各个领域的广泛应用，越来越多的学者开始将机器学习、深度学习用于股票市场的精准预测和分析，该方法的应用使研究者能够通过使用现有的数据来预测金融市场，使用新技术预测股票趋势成为了研究热点。机器学习作为一种计算机程序，可以从过去的数据中识别模式，对现有数据中学习，并产生预期的结果，从而推断股票价格。机器学习的方法不仅能够预测股票价格，而且还有助于预测市场行为，甚至提供新的模型来帮助处理数据源和预测市场波动。Patel 等人[21] 使用四种机器学习模型：人工神经网络 (Artificial Neural Networks, ANN)、SVM、随机森林 (Random Forest, RF) 和朴素贝叶斯 (Naive Bayes, NB) 预测股票市场价格走势，并将四种方法的性能进行了比较。Lu 等人[22] 提出了一种基于卷积神经网络 (Convolutional Neural Networks, CNN) 以及 LSTM 的股票价格预测方法，作者采用 CNN 从数据中高效提取特征，然后，使用 LSTM 对提取的特征数据进行股价预测。随着深度学习技术的进步，越来越多的研究人员开始尝试使用深度神经网络来预测股票市场情况，效果相比于传统的机器学习算法有明显提升。

深度学习方法通过设计集成的网络结构从原始股票数据集中提取股票数据特征，可以检测和利用现有金融经济理论不可见的数据模式。Long 等人[23] 基于神经网络提出了一种新型的端到端模型，将卷积神经元和循环神经元集成在一起，构建多过滤器结构，用于对股票样本的特征提取进而预测股票波动。Singh 等人[24] 在纳斯达克的谷歌股价多媒体数据集上证明了深度学习可以提高股票市场预测的准确性。Nosratabadi 等人[25] 统计分析发现 LSTM、CNN 以及 RNN 是股票预测模型中最常用到的三个神经网络结构。AkitA[26] 等人通过段落向量将报纸内容转化为其分布式表示，并使用 LSTM 模型挖掘了过去发生事件对股票价格的影响。Yadav 等人[27] 为印度股票市场开发了一个优化的 LSTM 模型用以预测股市的时间序列，Rather 等人[28] 使用自回归移动平均模型和指数平滑模型以及 RNN 来预测股票收益。Rasheed 等人[29] 使用一维卷积神经网络 (1D-CNN) 和 LSTM 组建模型预测股票价格。大量的研究表明，在利用时间序列类数据对股票趋势进行预测的应用上，深度学习模型表现良好。

1.2.2 股票非结构化数据源研究概述

从股票市场出现开始，股票预测因能够给股票投资者带来超额收益而吸引了大量的研究者进行探索。近些年来，互联网时代股票的泡沫化，使得金融分析师和经济专家们所推荐的传统股票估值工具不再受到青睐，股民们更倾向于通过其他方式获取信息来源来指导投资决策。在研究数据上突破了早期研究中仅利用股票结构化数据作为输入变量的股价趋势预测。互联网信息的爆炸式增长促进了社交新闻的发展，一些重大的新闻事件、社会舆论以及政府政策等成为影响股票投资者决策行为的重要因素。此外，大量的股票投资者在股票相关社交平台上探讨对于股票市场的相关看法，产生了具有极大研究价值的股票信息。同时，互联网与人工智能算法的发展为股票趋势预测带来了全新的研究角度与技术手段。这些突破传统结构化数据的信息来源以及崭新的研究技术为研究者探索股票市场的波动提供了新的思路。近年来，大量学者进行了许多尝试来挖掘在线信息来对股票市场进行预测。Li 等人[30] 通过将股票市场的结构化信息与社交媒体的非结构化信息相结合，使用两级信息融合方法来研究社交媒体文本的参与对股票价格的影响。Li 等人[31] 从情绪角度分析了社交新闻与股市的关系，作者构建了一个情感词典，然后将新闻文本投射到情绪词典中从而研究股票市场。Liu 等人[32] 通过量化事件的不同影响来提高预测的准确性，首先通过权重分配过程量化多个新闻和社交媒体中包含的有价值信息的重要性。然后通过其在隐藏层中的向量表

示，从事件中学习更多的上下文信息。上述研究充分证明，通过对社交平台中股票非结构化文本进行深层分析从而预测股票价格未来涨跌趋势是一种行之有效的方法。

尽管在利用非结构化数据进行股票预测方面已经有了很大进展，然而，在分析利用和目标股票相关的众多新闻时，等大多数研究工作都将每一条信息看的同等重要。Li 等人[33] 回顾了 229 篇关于量化网络媒体和股票市场之间相互作用的研究文章，这些文章来自金融、管理信息系统和计算机科学领域。作者根据媒体类型对代表性作品进行分类，但并未对这些文章的重要性进行区分。然而，真正的股票市中不同的文本信息对于股票市场的波动具有不同程度的影响，因为一些新闻显然包含重要的股票趋势信息，但另一部分新闻与股票趋势的相关性较低。在实际应用中，股票时序数据源的质量高低对股票预测的准确性会产生非常大的影响。Adam 等人[34] 表明股票数据的高噪声会导致股票准确预测的难度加大。新闻报道等文本数据具有的非结构化性质和高噪声性使它们难以直接作为股票预测模型的输入，数据源的高效性将会显著的提高股票预测精度和效率。为了从这些高噪声、低质量的股票文本中获取到更加精确的语义信息，NLP 技术以及文本分析等方法开始被研究者们广泛应用于处理股票相关的非结构化信息。这些技术能够从海量且杂乱的在线文本和社交媒体数据中提取股票相关的深层信息。Xing 等人[35] 对基于 NLP 技术的预测金融文章进行了综述，列出了用作股票预测输入的金融文本的类型以及它们的处理方式。此外，还描述了建模和实现细节中涉及的算法。Nassirtoussi 等人[36] 根据文本输入的性质以及市场类别，预处理程序和建模技术类型对文章进行分类。社交媒体是比公共新闻更加具有时间敏感性的信息源，因此，最近使用诸如推文之类的文本来预测股票走势成为了新的热点，在此基础上更深层次的探索这些文本的特征信息也受到了相当大的关注。Araci 等人[37] 提出了有关 NLP 的模型 FinBERT，使用大型金融语料库进行微调，分析金融文本所包含的情绪。Liu 等人[38] 使用 Transformer 编码器对社交媒体内容进行深度挖掘，提取其更深层次的语义特征，然后通过胶囊网络捕获文本的结构关系。因此，文本数据的质量以及可信度应当引起重视，对重要的文本信息增加关注以提高股票预测的性能。注意力机制是受到人类生物系统的启发，在处理大量信息时，人们往往专注于重要的部分，随着深度神经网络的发展，注意力机制在很多领域得到了广泛应用。最近，Hu 等人[39] 根据人类在面对高噪声、低质量、混乱新闻时的学习过程，基于顺序的内容依赖、多样化影响、有效和高效的学习三个原则，尝试从混沌的新闻数据中挖掘有效的市场信息进行股票预测，采用新闻层注意力机制和时间层注意力机制对低质量的新闻文本进行处理，然后再利用新闻文本进行股票趋势预测任务。不足的是，作者并没

有对单词层面存在的噪声进行处理，导致股票文本的语义信息表达不清楚，降低了股票趋势预测的准确性及可靠性。此外，作者只是使用推特文本作为影响股票价格的因素进行预测，并没有考虑历史交易数据的影响。

在以上关于股票预测的研究中，少数研究将非结构化数据与历史股票价格相结合，实现股票市场的预测。Huynh 等人[40] 探索了在线财经新闻和历史股票价格数据对未来的股票市场趋势的影响，并使用双向门控循环单元 (Bidirectional Gated Recurrent Unit, BiGRU) 作为预测模型。实验结果表明该模型简单但非常有效，与单纯使用历史价格信息的其他方法相比，可以显著提高股票预测精度。Xu 等人[41] 提出了一种基于推特文本和历史价格数据的股票预测生成模型，并取得了良好的效果。对比实验表明，使用社交媒体上收集的推文与实际股价数据相结合作为数据源对股票运动进行预测，这种组合方法优于单独分析股票价格或推文。主要原因是社交媒体评论和股市实际价格都是与股市密切相关的因素，直接反映了股票走势。每一种与股票市场相关联的数据都可为股票预测目标任务提供一定的信息，多源信息融合的目的是最大限度的对多种来源的数据进行综合分析、判断以便更好地完成目标。因此，在解决时间依赖问题的基础上，对非结构化文本和股票交易结构化特征进行针对性的特征抽取进行结合，从而创造出更有价值的新信息，有助于从根本上提高股票预测的准确性。

社交媒体中的股民情绪以及相关股票非结构化文本数据中包含的情感都会对股价产生影响。为了深度挖掘推文数据中隐含的股票情绪信息，Tetlock 等人[42] 首次研究了金融新闻文件中负面词语对股票市场的作用，作者探讨了媒体内容中的“悲观”指数与道琼斯工业平均指数之间的关系。“悲观主义”最基本的形式是《华尔街日报》上的负面词语数量。推特中的评论帖子的宏观层面情绪及其与总体市场指数的关系，吸引了大量学者的关注。Bollen 等人[43] 提出了两种情绪分析模型追踪和分析每日推特中重要的文本内容，以捕捉和分析公众情绪的变化。其中，OpinionFinder 将人的情绪分为积极和消极两种模式，而 GPOMS 将情绪进一步细分为六种类型，分别是平静、警觉、确定、重要、善良和快乐。实验结果表明，通过分析每日推特帖子的某些总体“情绪”得分，将其纳入特定的公众情绪类别，可以有效提高道琼斯工业平均指数预测的准确性。在此之后，大量的工作在研究标准普尔 500 指数和道琼斯平均指数中哪些个股更接近推特情绪预测[44, 45]。然而当前仍然缺乏预测个股收益变化的工作，因此本文对不同行业中的多支个股进行预测，并通过实验量化股价，进而验证情感分析对股票预测模型的增强作用。金融市场每秒产生大量公开数据，这些数据联系复杂很难

在单纯的经济模型中解释或推测。深度学习层次结构具有扩展输入数据以包括所有可能相关的特征的优势。将深度学习模型应用于这些问题可以产生比传统方法更健壮和有用的结果。Heaton 等人[46] 介绍了各种深度学习分层模型在财务预测问题中的应用，包括用于模型选择的 dropout、自编码器和 LSTM 模型。受到 Vaswani 等人[47] 提出的 Transformer 模型在自然语言处理中对顺序数据进行建模的成功启发，本文将其作为预测股票收盘价的预测模型。与传统的深度学习框架相比，Tranformer 中的自注意力机制可以并行训练，从而更容易获得全局信息。

1.2.3 不确定性量化研究概述

深度学习模型在股票预测任务中取得了巨大的成功。然而尽管相关模型具有出色的预测性能，但是很容易在预测中产生过度自信的情况。Nguyen 等人[48] 发现了一个很有趣的现象：深度神经网络很容易被欺骗，对于无法识别的样本模型看作是具有高置信度的可识别对象，也就是说深度神经网络模型对于与训练数据差别较大的样本会做出错误但非常有信心的预测，这些错误的预测在实际应用中可能会产生高昂的代价。为了解决这一问题，文章提出通过不确定性量化来使深度学习网络知道自己不知道的能力。近些年来，不确定性量化被广泛应用于众多领域。2020 年，Chang 等人[49] 将数据不确定性学习应用于人脸识别，尝试在人脸识别系统中捕获数据不确定性，通过学习特征和不确定性进而提升人脸识别模型的性能。齐现英等人[50] 提出一种基于不确定性信息融合的中智灰滤波算法，对图像噪声的不确定性进行量化，并利用不确定性融合信息对图像噪声进行降噪。Dolezal 等人[51] 将不确定量化用于癌症数字组织病理学领域，使用 dropout 估计不确定性然后计算训练数据的阈值，以建立低置信和高置信预测的界限。最后训练模型来识别肺腺癌和鳞状细胞癌。Abdar 等学者[52] 研究回顾了深度学习中使用的不确定性量化方法的最新进展，研究了这些方法在机器学习、深度学习以及强化学习中的应用，并强调了与不确定性量化领域相关的基本研究，为相关领域的研究人员提供文献综述。

常用的不确定性量化方法包括集成方法、蒙特卡洛方法 (MC dropout) 和贝叶斯神经网络等。Malinin 和 Gales[53] 在一个统一的、可解释的基于概率集成的框架内研究自回归结构化预测任务的不确定性估计，考虑序列数据在令牌级和完全序列级的不确定性估计、各种不确定性量化的解释和应用，并讨论相关的理论和实践挑战。Gal 和 Ghahramani 等人[54] 提出了使用 MC dropout 来估计不确定性。MC dropout 并不需要

改变模型训练方式，只需在预测时也同样开启 dropout，这样在 dropout 的影响下，可以得到不同的输出结果，求得这些输出结果的平均值和方差，平均值即为模型的预测结果，方差即为模型的不确定性。贝叶斯神经网络[55] 将神经网络的参数看做服从一定的先验分布，给定输入时可以产生具有一定分布性质的输出用来估计模型的不确定性。贝叶斯模型虽然可以提供对模型不确定性进行推理的能力，但是这种方法实际上很难应用到大的网络结构中，因为此方法的训练难度较大，而且其指定先验参数非常困难。因此，越来越多的研究人员考虑非贝叶斯的方法来量化模型的不确定性。Wu 等人[56] 为分类任务引入了一种基于神经网络的 logit 输出的不确定性量化方法。文章提出，相同类别样本的 logit 输出应该相似。因此可以根据训练过程中预测正确样本的 logit 向量与样本输出的 logit 之间的差距确定不确定性。相比于其他不确定性度量方法，logit 不确定性方法与学习过程以及培训任务没有关联，且在处理高维小样本、非线性等预测方面具有良好的性能。此外，由于该方法适合于任何产生 logit 输出的网络模型，相比于其他不确定性方法，该度量方法比较灵活，伸缩性强，具有通用性，相比于贝叶斯方法，该方法避免了指定模型先验分布和推断后验分布的需要。Kong[57] 等人构建了一种新的量化深度学习网络不确定性的模型 SDE-NET。其他方法相比，该模型同时考虑了偶然不确定性和认知不确定性并分别进行建模，并且能够在其预测中分离出不确定性的两个来源。

现有的股票预测任务对不确定性因素缺少研究。针对股票数据中的低效率、低质量问题造成的预测效果不理想这一现象，可以考虑从不确定性量化方面入手，将训练数据进行不确定性量化，选取低不确定性值的数据进行分析预测，过滤掉高不确定性值的数据，进而减少误差，提供预测性能。

1.3 研究内容和创新点

基于当前的股票趋势预测模型方面存在的问题，本文将基于深度学习对以下内容进行研究：其一，由于现有的来自社交媒体的文本型数据量巨大，质量参差不齐，考虑到现有的股票预测模型的研究并没有对单词层面存在的噪声进行处理，导致股票文本的语义信息表达不清楚，降低了股票趋势预测的准确性及可靠性。本文使用三层注意力机制模型，通过单词注意力层、句子注意力层以及时间层注意力层对股票文本数据进行处理。而且传统的股票预测大多只是基于股票价格数据或者社交文本数据，或者对二者进行简单融合进行预测，缺乏对股票异构数据的分析处理，因此，本文提出

了一种基于异构数据和多层注意力机制的股票预测网络。解决了以往研究股票时数据源单一和未考虑股票文本对股票波动趋势影响的问题。其二，本文提出了基于特征选择和不确定性量化的股票预测网络，使用确定性增强股票预测性能。首先，对股票价格指标特征进行选择，提高股票模型效率与准确性。在非结构化数据方面从不确定性角度切入，通过对股票输入数据中的不确定性进行量化，选取不确定性较低的数据参与股票的预测过程，提高输入可用性，通过增强股票数据的可靠性提高股票的预测性能。最后使用能够区分不同不确定性类型的 SDENet 网络作为预测模型，减少误差，获得收益。其三，本文提出了基于情感分析和 Transformer 模型的股票预测网络，将股票投资者情绪纳入到股票价格特征中，作为后续模型的输入源。然后使用改进后的 Transformer 模型增强股票预测模型性能。然后在多个不同行业的个股数据集中进行实验，验证提出模型的稳定性和普适性。

在基于异构数据和多层注意力机制的股票预测网络中，主要完成以下研究工作：

(1) 本文使用 BiLSTM 网络捕捉股票历史价格交易数据中的长期依赖关系，解决了传统基于股票价格进行预测的模型中未考虑时间因素或者仅考虑时间上单向依赖关系的问题。

(2) 为了更好的处理股票文本特征中干扰数据过多、数据可用性不高等问题，本文使用两层注意力机制对股票相关信息进行处理，除了在新闻层加入了注意力机制，在股票文本数据的单词层也引入了注意力层。相比于之前的深度神经网络，加入单词层注意力机制的模型结果更加高效、准确。

(3) 在上面提到的双重思想的启发下，本文提出了一种新的股票预测网络 CredibleNet：将经过处理后的历史价格数据和新闻文本数据进行特征融合，然后使用时间层注意力机制获得高质量的股票数据，使用 LSTM 进行股票趋势预测。

在基于特征选择和不确定性量化的股票预测网络中，主要完成以下研究工作：

(1) 为了解决价格特征中的冗余问题，提高预测模型的精确度，本文使用极端决策树算法对股票的历史价格特征进行处理，选取对股票行情高度相关的特征简化模型。

(2) 针对股票相关数据中包含的低质量不可信数据，提出了一种基于不确定性量化的股票预测方法，使用混合高斯模型对正确预测的 logit 度量数据进行建模，进而获得不确定性量化数值，过滤高不确定性值的不可靠数据，保留不确定性较低的数据对股票趋势进行预测。

(3) 以能够获得良好预测精度的 drift net 和描述随机环境中模型不确定性的 diffu-

sion net 构成的 SDENet 网络作为股票预测模型，该模型能够从具有噪声的大量股票数据中选取低不确定性数据进行预测。

在基于情感分析和 Transformer 模型的股票预测网络 SenTransNet 中，主要完成以下研究工作：

(1) 在股票文本数据上进行情绪分析，将股票技术指标与市场中的公众情绪相关联，进而增强预测模型性能。

(2) 提出一个基于推文情绪和历史股价数据的 Transformer 模型来预测股价的未来走势，以股票文本情感分析以及价格指标的融合数据作为输入，以改进后的 Transformer 模型为预测模型，构建股票预测网络 SenTransNet。

(3) 针对上述模型，使用来自不同行业的多支股票数据集进行测试，验证所提出模型的预测性能以及鲁棒性。

1.4 论文组织结构

本文各主要章节的安排如下：

第一章：绪论。本章节主要介绍了股票预测任务的现有背景和研究意义，并且对现有的各种股票预测方法进行综述，归纳国内外目前的股票预测成果。通过对现有的股票预测方法存在的问题展开研究，确定了本文的研究目的并对论文的研究内容以及创新点进行了总结。最后对本论文的各章安排进行简要说明。

第二章：相关技术介绍。本章对股票预测研究中所涉及到的相关理论和技术进行了详细概述。包括长短时记忆网络、双向长短时记忆网络、注意力机制、决策树、基于 logit 不确定性量化方法以及 Transformer 模型。

第三章：基于异构数据和多层注意力机制的股票预测网络 CredibleNet。本章首先对股票预测进行任务介绍，然后对模型框架进行描述，包括单词层和句子层注意力机制去除低质量文本信息、特征融合股票异构数据、时间层注意力机制捕获高效融合数据，将处理后的数据输入到 LSTM 进行股票趋势预测。最后进行实验，消融实验部分说明所提出模型的不同模块对股票趋势预测的影响，对比实验比较不同方法在同一数据集中的指标情况，分析实验结果以及模型表现，并给出所提出模型的不确定性量化分析。

第四章：基于特征选择和不确定性量化的股票预测网络 LogNet。本章首先对使用的不确定性量化的方法进行描述，介绍了基于 logit 不确定性量化的可用性。然后使

用 SDENet 网络作为股票趋势的预测网络，该网络能够从认知不确定性和偶然不确定性两方面考虑，选取低不确定性的股票数据，提高模型的可靠性和准确性，为模型应用于现实股票市场提供保障。最后根据评价指标，对模型和现有算法进行对比实验，得到实验结果并进行分析。

第五章：基于情感分析和 Transformer 的股票预测网络 SenTransNet。本章首先使用 LM(Loughran and McDonald, LM) 金融情感词库提取推文中的文本情感特征，然后将其与股票价格特征进行特征融合，将融合后的数据输入到改进后的 Transformer 模型中进行预测输出。然后在多支不同行业的数据集上进行实验分析，验证所提出模型的有效性和鲁棒性。

第六章：总结和展望。本章对提出的基于异构数据和多层注意力机制的股票预测模型、基于特征选择和不确定性量化的股票预测网络和基于情感分析和 Transformer 的股票预测方法研究进行梳理总结。此外，提出了目前研究工作中的不足，对股票趋势预测未来可能的发展趋势进行展望。

第 2 章　相关技术介绍

2.1　文本表示方法

对于影响股票趋势的的非结构化因素，计算机无法直接理解这些文本的含义，这些文本中最小组成的是单词，单词组成语句，语句再构成文档。因此处理这些非结构化数据，首先要从单词入手。词嵌入 (Word Embedding) 表示方法通过将文本语料库中的某个单词映射到数值向量空间，将文本数据转换成数值型向量。从而使得模型能够分析文本中隐藏的信息。

2.1.1　Wordvec

2013 年，谷歌的 Mikolov 等人提出了一种将文本数据转换为数学向量的词嵌入方法 Word2vec[58]。其通过在大规模文本语料库上对词向量训练，生成包含单词之间前后联系的具有语义表征能力的词向量。该方法在自然语言处理领域中有广泛的应用，例如文本分类、词汇相似度计算、语言翻译等。Word2vec 通过学习词汇的分布式表示，能够捕捉单词之间的语义和语法关系转化为数字向量。使得计算机更好地理解和处理自然语言数据。其核心思想是将每个单词表示为一个固定长度的向量，然后通过一个预测目标函数学习这些向量的参数。该方法的训练模型是包含一个隐藏层的前馈神经网络，网络的输入和输出均为一个基于独热编码 (One-Hot) 的向量，通过两个参数矩阵来训练最终的词向量结果。如图2.1所示，所有样本基于滑动窗口进行上下文构造，当训练神经网络使用所有样本，最终获得输入层到隐藏层训练的权重矩阵，每一个 One-Hot 编码代表其中的一行作为词向量。这样就可以将原始的 V 维向量转化为 N 维向量，并在词向量之间保持一定的相关性。

基于词向量的训练模式有两种训练模型：连续词袋模型 (Continuous Bags-Of-Words, CBOW) 和跳字模型 (Skip-gram)。CBOW 的主要思想是通过通过周围的单词预测中心目标单词以及它的意思，它属于基于神经网络的词嵌入方法之一。CBOW 模型首先将上下文中的每个单词表示为一个单词嵌入向量，并将这些向量相加以得到一个上下文向量。接下来，这个上下文向量被传递到隐藏层，隐藏层将上下文向量进行线性变换，并通过非线性激活函数 (如 ReLU) 将其激活，以生成一个隐藏层的输出向量。最后，输出向量被传递到输出层，并通过 softmax 函数将其转换为概率分布，从

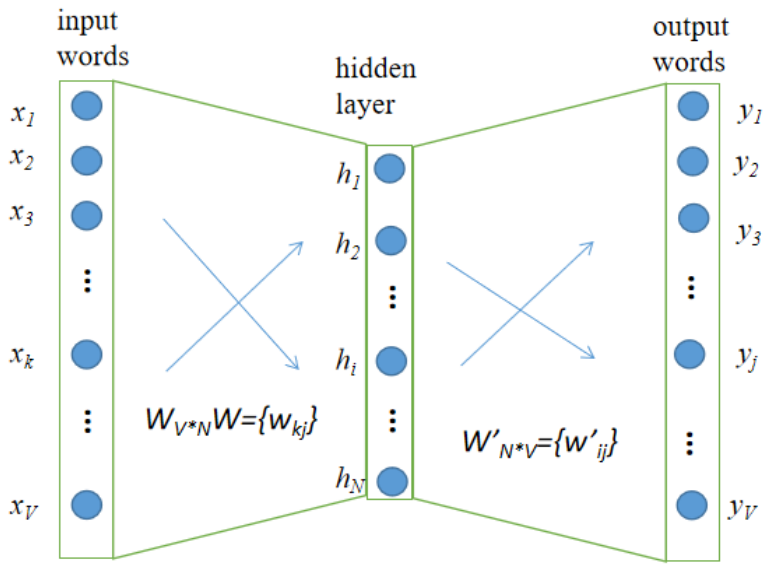

图 2.1 Word2vec 训练模型

而预测目标单词。CBOW 模型的优点是它可以在小型数据集上进行训练，并且可以生成高质量的单词嵌入。由于它是一种无监督的学习方法，因此它可以利用大量未标记的文本数据来生成单词嵌入，这对于文本处理非常有效。

Skip-gram 模型的目标是通过给定一个单词来预测其周围的上下文单词。与 CBOW 模型不同，Skip-gram 模型的输出是多个单词嵌入向量，而不是单个单词嵌入向量。这些单词嵌入向量表示与输入单词相关的上下文单词。Skip-gram 模型首先将每个单词表示为一个向量，这些向量被初始化为随机值。然后，对于每个单词，模型会从文本中选取一些上下文单词，并将这些上下文单词作为输出。对于每个上下文单词，Skip-gram 模型都会使用输入单词的嵌入向量来预测它的嵌入向量。这个过程可以通过一个简单的神经网络来实现，该神经网络由一个输入层和一个输出层组成。输入层包含输入单词的嵌入向量，而输出层包含上下文单词的嵌入向量。中间的隐藏层通常是一个全连接层，可以通过非线性激活函数来增加模型的表达能力。Skip-gram 模型的一个优点是它可以学习到非常高质量的单词嵌入，这对于许多自然语言处理任务非常有用。与 CBOW 模型相比，Skip-gram 模型通常需要更长的训练时间和更大的数据集，但可以产生更准确的单词嵌入。图2.2和图2.3分别为连续词袋模型和跳字模型结构图。

在无监督的情况下，上述方法的预测精度与单词嵌入的维度有关。增加嵌入维度可以提高预测准确度，直到达到最佳嵌入维度的收敛点，最终得到一个 g 在不影响

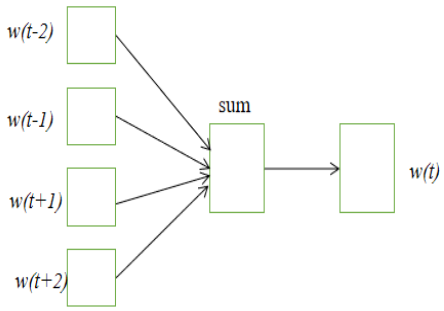

图 2.2 连续词袋模型结构图

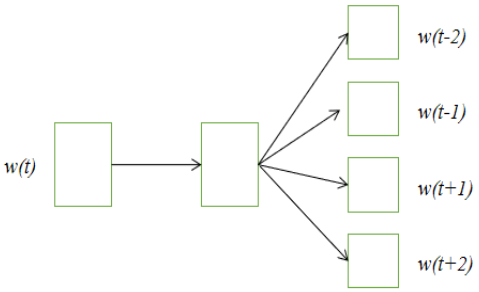

图 2.3 跳字模型结构图

准确度的情况下具有最小维数的取值。在 CBOW 中，假设目标词的上下文的顺序不对预测结果产生影响，采用平均上下文来进行加权。而在 Skip-gram 模型中，每一个上下文向量都被独立地进行加权和比较，相邻的上下文单词比距离较远的单词权重更大。虽然通过 Word2vec 获取的词间关系简单直观，表示词向量可靠有效，但是传统的词嵌入算法为每个词分配相同的向量，无法解决一词多义问题。

2.1.2 Glove

CBOW 和 Skip-gram 模型方法具有一定的局限性，这两个模型都是基于上下文信息来学习单词嵌入的。因此，在处理一些没有上下文信息的场景时，它们可能不是最佳选择。其次，这两个模型都是基于局部上下文来进行单词嵌入的，这意味着它们可能无法捕捉全局的语义信息。当处理一些复杂的语义关系时，这些模型可能无法提供最佳的单词嵌入。此外，这两个模型在对稀有单词的处理、对多义词的处理、对长文本的处理等方面也存在缺陷。

为了解决这些局限性，2014，Pennington 等人[59] 提出了一种基于全局的将单词映射到向量空间中的词向量学习算法，即 Glove(Global Vectors for Word Representation, Glove)，它的目标是学习出单词之间的语义关系，使得在向量空间中距离相近的单词具有相似的含义。Glove 算法的核心思想是将语料库中单词出现的统计 (共现矩阵) 看作学习词向量表示的关键，利用了全局语料库的统计信息生成单词向量表示。具体来说，它通过分析大规模文本语料库中单词的共现矩阵来进行训练。共现矩阵是一个记录了每对单词在同一个上下文中出现的次数的矩阵，其中上下文可以是单词周围的若干个单词或者整个文档。Glove 算法首先通过共现矩阵计算出每对单词之间的相关性，并将这些相关性作为目标函数。然后，通过最小化这个目标函数来学习每个单词的向量表示。Glove 算法的优点是能够同时考虑到全局的共现信息和局部的语义信息，从

而得到更准确的单词向量表示。使用 Glove 算法得到的单词向量可以应用于自然语言处理中的各种任务，例如词汇相似度计算、情感分析、语言模型等。Glove 算法已经被证明在许多任务上比传统的基于词袋模型的方法表现更好。

从本质上说，Glove 是具有加权最小二乘法目标的对数双线性模型。字词共现概率的比率又编码成某种形式的潜在可能意义。表2.1是基于 60 亿词汇语料库关于冰和蒸汽的词的共现概率，显示了在大语料库中的这些概率值，以及彼此间的比值。与原始概率相比，比值更能区分相关的单词 (比如 solid 与 gas) 和不相关的单词 (water 与 fashion)，并且也能更好地区分这两个相关的单词。

假设，X 为单词-单词的词频共现矩阵。其中的 X_{ij} 表示单词 j 出现在词语 i 上下文的频数；并且令 $X_i = \sum X_{ik}$ 为任意单词出现在单词 i 上下文的次数之和；令 $P_{ij} = P(j|i) - \frac{X_{ij}}{X_i}$ 为单词 j 出现在单词 i 上下文的概率。令 $i = ice$(冰) 和 $j = steam$(蒸汽)。这两个单词之间的关系能通过研究它们基于不同单词 k 得到的共现矩阵概率之比得到。对于与 ice 有关但与 $steam$ 无关的单词，比如说 $k = solid$(固体)，比值应该会很大。同样，对于与 $steam$ 有关系然而与 ice 无关的词语，比如 $k = gas$(气体)，比值应该很小；而对于那些与两者都相关或都不相关的单词，比如 $water$(水) 或 $fashion$(时尚)，比值应该接近于 1。

Glove 的训练目标是通过学习单词的共现矩阵来得到单词的向量表示，使得在向量空间中距离相近的单词具有相似的语义含义。即最小化以下损失函数：

$$J = \sum_{i,j=1}^{V} f(P_{ij})(w_i^T \hat{w}_j - b_{ij} - logP_(ij))^2 \tag{2.1}$$

其中，V 表示词汇表的大小，P_{ij} 表示单词 i 和单词 j 在同一上下文中出现的概率，w_i 和 $\hat{w}_j$ 分别表示单词 i 和单词 j 的向量表示，b_{ij} 是一个偏差项，$f(P_{ij})$ 是一个权重函数，用于加权不同的共现信息。

通过最小化上述损失函数，Glove 可以学习到在向量空间中具有相似语义含义的单词的向量表示。在本文中，选用 Glove 将推文转化为词向量，然后进行后续操作。

2.1.3 LSTM

时间依赖对于股票运动预测依旧非常重要，因为更接近目标交易日的数据对预测结果的影响更大。股票预测具有高度非线性，加上股票市场具有时间序列的特性，因

表 2.1 冰和蒸汽的词的共现概率

Probability and Ratio	k=solid	k=gas	k=water	k=fashion
P(klice)	1.9*10-4	6.6*10-5	3.0*10-3	1.7*10-5
P(klsteam)	2.2*10-5	7.8*10-4	2.2*10-3	1.8*10-5
P(klice)/P(klsteam)	8.9	8.5*10-2	1.36	0.96

此 RNN 是一种适合用于股票预测的模型。然而，尽管 RNN 允许信息的持久化，但它在处理具有长记忆性的时间序列数据方面的表现不够强大。当面对大量数据或者过长的序列数据时，RNN 可能会遭受梯度爆炸或消失的问题，导致训练变得异常困难。当数据过多过长时，该模型会遗忘之前的数据，导致产生测量误差。LSTM 在 RNN 的基础上进行了改进，有效地解决了其在反向传播过程中存在的数据遗忘和梯度爆炸等问题。对于股票数据而言，LSTM 能够有效的解决股票预测模型中股票数据过长、内容过多的问题。

LSTM 是一种特殊的循环神经网络模型，能够有效的处理长记忆性问题，它在 RNN 的基础上加入了存储单元和逻辑门结构。存储单元 c_t 用于记录神经元的状态，门结构可以捕捉输入特征中的长期相关信息。LSTM 模块结构图如图2.4所示，x 表示输入向量，h 代表模型的输出向量，C 为单元状态，f，i，o 分别代表遗忘门、输入门、输出门的激活值。

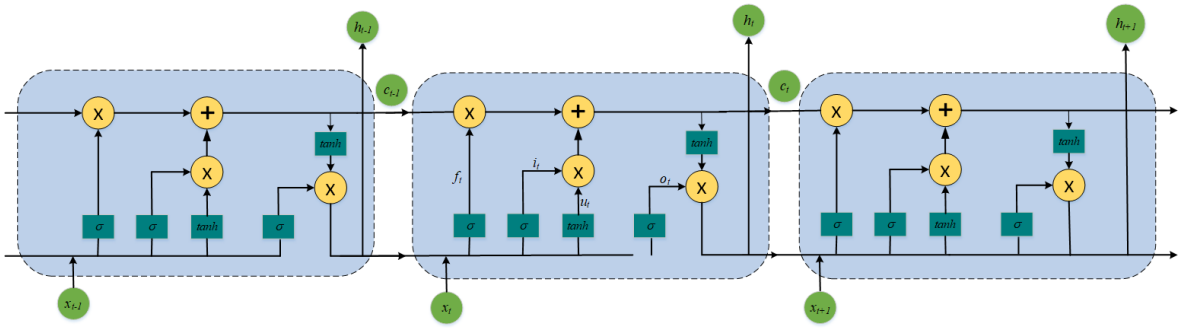

图 2.4 LSTM 神经网络结构

在时刻 t，遗忘门的激活值 f_t 计算为：

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{2.2}$$

f_t 是由输入值 x_t、$t-1$ 时的输入 h_{t-1} 以及遗忘门的偏差矩阵 b_f 共同获得，然后通过 sigmoid 函数进行归一化处理，遗忘门决定了从之前的状态单元 C_{t-1} 中丢弃多少信

息。然后，通过计算输入门的激活值 i_t 以及当前状态候选值，决定新的状态单元 C_t：

$$\overline{C_t} = \tanh\left(w_C\left[h_{t-1}, x_t\right] + b_C\right) \tag{2.3}$$

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

$$o_t = \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{2.5}$$

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

从而，可以根据 o_t 的和 C_t 的值得到输出值 h_t 的值：

$$h_t = o_t \tanh\left(C_t\right) \tag{2.7}$$

2.2 双向长短时记忆网络

BiLSTM 是在 LSTM 的基础上发展而来的一种神经网络模型，由两个独立的 LSTM 构成，输入数据分别以正向和逆向的顺序输入到两个 LSTM 神经网络中，将两个输出向量进行拼接融合后作为最终特征表达，两层网络协同发挥作用对结果产生影响。与传统单向的 LSTM 相比，BiLSTM 利用双向轨迹序列信息进行输入，正向 LSTM 可以捕捉到输入序列中前面的信息，而反向 LSTM 则可以捕捉到输入序列中后面的信息。这使得该模型能够更好地处理时间序列数据，有利于提高模型预测的性能。股票预测任务受到历史和未来多个输入共同影响，尤其在趋势波动较大时，前后数据差别比较大，所以使用 BiLSTM 捕获历史和未来时刻的数据信息可以得到更精确预测结果，其结构如图2.5所示。箭头表示信息流动的方向，变量 x_1、x_2、x_3、x_4 为输入数据，y_1、y_2、y_3、y_4 为相应时刻的最终输出，L_1、L_2、L_3、L_4 与 B_1、B_2、B_3、B_4 分别表示 LSTM 隐藏状态在不同时间的前向和后向迭代，w_1、w_2、w_3、w_4、w_5、w_6、w_7 代表参与每一层的计算的权重值。自前向后 LSTM 的隐藏更新状态 L_i:

$$L_i = f((x_i w_1) + (w_2 L_{i+1})) \tag{2.8}$$

自后向前的 LSTM 的隐藏更新状态 B_i：

$$B_i = g((x_i w_3) + (w_5 B_{i-1})) \tag{2.9}$$

最终输出结果 y_i：

$$y_i = h((L_i w_4) + (w_6 B_i)) \tag{2.10}$$

上式中，$f(x)$、$g(x)$、$h(x)$ 为不同隐含层中使用的激活函数。

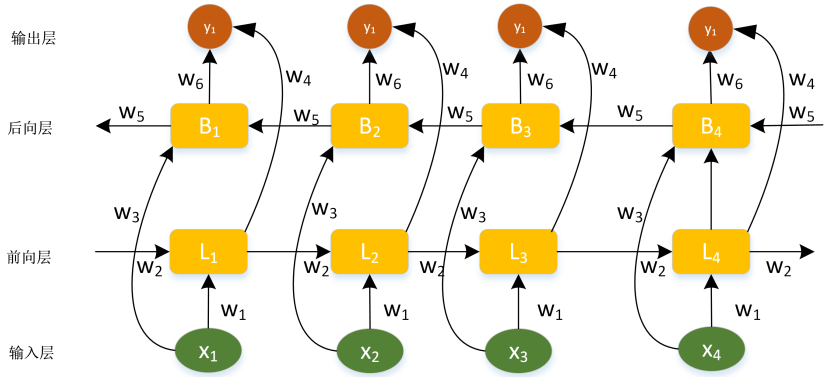

图 2.5 BiLSTM 神经网络结构

根据上述描述可以看出，BiLSTM 的特殊结构使其能够对输入的时间序列数据进行前后两个方向的计算，有助于模型学习前后时间数据序列之间的相关信息，提高模型的预测精度。

2.3 注意力机制

注意力机制作为一种新的计算机制首先应用于机器视觉领域。近年来，注意力机制开始广泛应用在股票预测方面，它的合理使用对股票预测模型的学习能力和预测精度起到了至关重要的作用。注意力机制是指让模型能够有选择的关注和处理具有不同重要性程度的信息，通过为关键特征分配更高的权重，使得模型更多的关注重要特征，提升模型预测性能。在股票的趋势预测模型中引入注意力机制可以增强输入数据关键特征的影响，提高模型预测效果。注意力机制通过预测数据各部分对预测目标的重要性程度，得到注意力权重，然后利用注意力权重增强数据中的重要部分并减弱数据中与目标关联性小的部分。注意力机制被广泛地应用在文本语言的处理和分析任务中，很大程度上提高了神经网络的非线性表达能力和深层语义的提取能力。

将互联网中存在的关于股票的文本数据特征，作为注意力模型的输入 x_i，注意力权重可以根据如下式 (2.11) 得到：

$$b_i = S(F(x_i)) \tag{2.11}$$

$S(.)$ 函数表示注意力模型对应的函数变换，用于学习输入数据 x_i 不同部分的重要程度，可选取 sigmoid 函数作为此学习函数。$F(.)$ 表示归一化操作，用于将注意力的权重限定到一定范围，可以选择 softmax、sigmoid、tanh 等函数实现。b_i 将得到的注意力权重 b_i 与原始特征进行对应元素相乘，从而完成对关键信息的选择，该过程公式如下:

$$n_i = b_i * x_i \tag{2.12}$$

其中，n_i 为注意力模型的输出，即对原始输入数据 x_i 加入权重之后的特征。使用注意力机制可以对输入的不同特征给定不同的权重，对原始数据的不同成分进行加强或减弱。特征注意力机制如图2.6所示，股票文本数据 x_i 作为特征注意力机制的数据输入，然后进行特征权重的计算后得到权重值 e_i，为了避免输入数据差异太大，再进行归一化操作后得到 b_i，并将注意力权重与对应特征做乘积从而得到相关输入信息的表达 $b_i x_i$

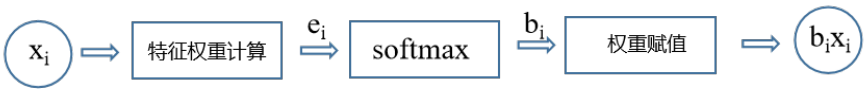

图 2.6 注意力机制流程图

2.4 极端随机树

如何选取历史价格数据中的特征也是研究者研究股票市场的一大难题。极端随机树 (Extremely Randomized Trees, Extra Trees) 是[60] 于 2006 年所提出的一种基于随机决策树的集成方法。集成学习通过将多个机器学习算法的预测结果进行集成，根据“少数服从多数”的原则做出预测完成学习任务，因而能够预测结果具有更高的准确性和鲁棒性。与随机森林类似，Extra Trees 也采用了随机化策略来构建多个决策树。然而，相比于随机森林，Extra Trees 采用了更加随机的特征选择方式和节点划分方式。在构建每个决策树时，其不仅随机选取特征子集，而且对每个特征随机选择一个切割阈值来进行节点划分。这种随机性使得该方法更加多样化，减少了过拟合的风险。在训练过程中，Extra Trees 通过对每个决策树的预测结果进行投票或平均来得到最终的预测结果。相对于单棵决策树，Extra Trees 集成了多棵决策树的预测结果，具有更高的鲁棒性和更好的泛化能力。

算法 2.1: 极端随机树拆分算法

Split_a_node(S) 输入：对应于要拆分的节点的学习子集 S

输出：[a<a_c] 或者空树

-如果 Stop_split(S) 返回值为 TRUE，返回空树

-否则在所有非常量候选属性 (S) 中选择 K 个属性a_1，... ，a_k

-绘制 k 个分支s_1，...，s_k，s_i=Pick_a_random_split(S，a_i)，$i = 1$，...，K

- 返回一个分支 $S*$，这样 Score($s*$，S)=$max_{i=1....k}Score(s_i\ S)$

Pick_a_random_split(S，a)

输入：子集 S 和属性 a

输出：切分位置

- 选择变量 a^s_{max} 和 a^s_{min} 表示 a 在 S 中的最大值和最小值

- 在 [a^s_{max}，a^s_{min}] 中画一个随机切点 a_c

- 返回切分位置 [$a < a_c$]

Stop_split(S)

输入：子集 S

输出：布尔值

- 如果 $|S|$<n_{min}，返回值为 TRUE

- 如果 S 中的所有属性都是常量，返回 TRUE

- 如果输出是在 S 中的常量，返回 TRUE

-否则，返回 FALSE

此外，Extra Trees 的随机性可以抑制过拟合，不会因为某几个极端的样本点而将整个模型带偏。将多种随机决策树聚集起来拟合到数据中，将多个弱学习器组合成强学习器。Extra Trees 的具体算法如算法 2.1 所示。该算法通过自上而下的方法构建决策树集合，它与其他决策树的最大不同是它通过完全随机选择左右分支来分割节点，并且它使用所有训练样本来构建决策树。相比于随机森林，极端随机树表现会更好，因为在随机的测试集中，某些特征仍然具有较高区分度，说明这个特征确实很重要。

对于极端随机决策树而言，特征个数 k 决定了特征选择过程的强度，n_{min} 决定了平均输出噪声的强度，M 决定了集成模型方差减少的强度。

2.5 基于 logit 的不确定性量化方法

不确定性描述了结果的可靠程度，如果不确定性值趋近于 1，表明预测结果完全不可靠，相反，如果不确定值靠近 0，则说明预测结果相当可靠。Wu 等人[56] 提出了一种基于神经网络输出的 logit 进行不确定性度量的方法。属于同一个类别的样本通过神经网络输出的 logit 向量应该具有相似的特性，如果预测类输出的 logit 值与已知样本的 logit 值差别很大，则说明预测可信度偏低。由此可以看出，基于 logit 的不确定性分析方法类似于估计 k 维欧几里得空间中的新数据点属于预测类的高斯混合的概率问题，具有高密度值的点应该具有较低的不确定性是有道理的。使用高斯混合分布函数 (GMMs) 对每一类中正确预测的样本的 logit 输出进行建模，然后基于高斯混合分布的概率密度函数 (pdf) 对预测结果的不确定性进行建模，概率密度函数的值越小，不确定性越大，意味着预测结果越不可靠。此外，为了能够概率密度的极小值，使用一个评分函数将概率密度函数转化为分数，然后使用 sigmoid 函数将评分值映射到 0 到 1 之间。

对于一个全连接神经网络，假设其隐藏层个数为 N，每个隐层大小为 $H_d(1 \leq d \leq N)$，输入输出向量的维数分别是 H_0 和 H_{N+1}，神经网络可以描述为：

$$g_i^{(u)}(x) = ReLU(f_i^{(u)}(x)) \tag{2.13}$$

$$b_i^{(u)} \sim N(0, C_b^{(u)}), u \in [1, ..., D] \tag{2.14}$$

则每个神经网络的输出极限分布收敛于高斯混合分布。$G(c)$ 代表具有 c 个分量的高斯混合分布。

2.6 Transformer 模型

Transformer 是 Ashish 等人[47] 提出的利用自注意力机制提高模型性能的网络结构。模型结构图如图2.7所示。Transformer 由编码器-解码器 (Encoder-Decoder) 两部分构成。编码器结构包括 6 层，每层由两个子层构成：一个子层是由多个 Self-Attention 组成的 Multi-Head Attention 层，另一个是 Feed Forward 层。每个子层都附加了一个 Add&Norm 层。

Self-Attention 层是 Transformer 中最核心的部分，通过将输入 X 进行线性变换得

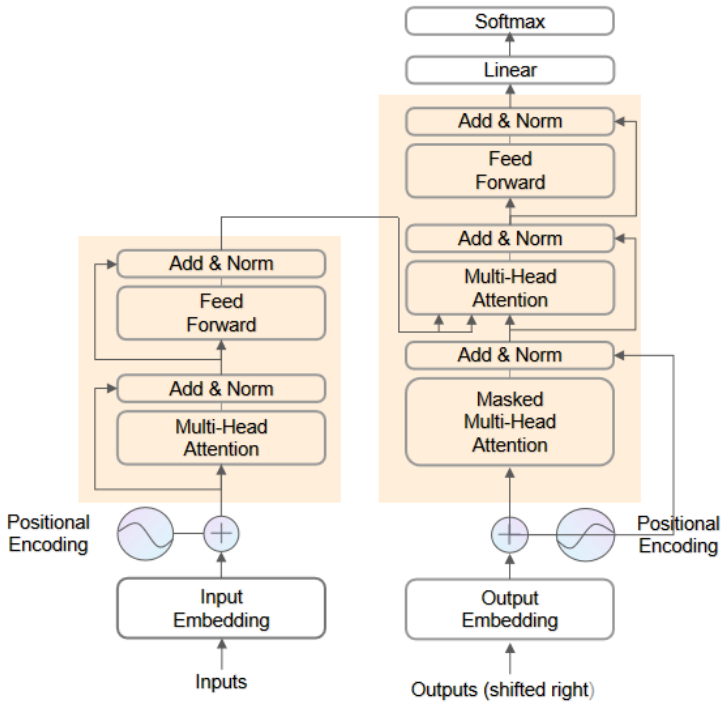

图 2.7 Transformer 模型结构图

到 Query 和 <Key，Value> 对，记为 Q，K，V，计算公式如下：

$$Q = X \cdot W_Q \tag{2.15}$$

$$K = X \cdot W_K \tag{2.16}$$

$$V = X \cdot W_V \tag{2.17}$$

Multi-Head Attention 由多个独立的 Self-Attention 集成。将多个 Self-Attention 得到的向量进行拼接形成一个大的特征矩阵，该特征矩阵经过全连接层后得到向量 Z。

解码器也有 6 层，每层包含三个子层：第一个子层是 Masked Multi-Head Attention 层，用于进行 Self-Attention，与编码器不同的是，由于解码器是序列生成过程，因此需要对 t 时刻之后的结果进行掩码处理；第二个子层是 Multi-Head Attention 层，用于计算 Encoder-Decoder 的注意力，这里不是 Self-Attention；第三个子层是 Feed Forward 层。与编码器相同，解码器的每个子层也包含 Add&Norm 层。

Transformer 模型的输入由输入嵌入层 (input embedding) 与位置嵌入层 (positional embedding) 相加得到。其中输入嵌入层对文本单词进行编码，将其转化为词嵌入向量，常用方法包括 Word2Vec、Glove 等。Transformer 模型的位置嵌入层用于捕捉单词间

的顺序信息。计算公式如下：

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d}}) \tag{2.18}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d}}) \tag{2.19}$$

Transformer 原结构被用于自然语言处理领域中机器翻译任务。在本文中将其进行改进后用做股票预测模型。

2.7 本章小结

本章对股票趋势预测问题中所使用的相关技术进行概述，包括长短时记忆网络、双向长短时记忆网络、注意力机制、极端随机树、基于 logit 的不确定性量化方法以及 Transformer 模型。本文将使用双向长短时记忆网络对雅虎财经获得的股票价格特征进行处理，使用双层注意力机制对股票相关的推文数据进行去噪、提取可靠特征。最后使用长短时记忆网络作为决策模型，预测未来股票趋势。并在此基础上建立股票预测网络 CredibleNet；使用极端随机树对股票的历史价格数据进行特征选择，利用基于 logit 的不确定性量化方法，通过消除具有高不确定性的数据、增强低不确定性数据的表达来提高预测模型的精度；将用于自然语言处理网络结构 Transformer 模型进行改进后引入到股票预测任务当中，提升模型的预测性能。

第 3 章　基于异构数据和多层注意力机制的股票价格预测模型

3.1　任务分析与描述

股票市场是一个极度混沌的系统，市场中的许多因素都会对股票未来的波动情况产生影响，例如股票的历史交易数据、行业发展状况、自然灾害、媒体报道以及市场情绪等。在之前的研究中，研究者大多于单一信息源如历史价格时间序列进行股票预测。为了从新闻文本中提取更深层次的潜在信息，本文将考虑多条新闻的综合影响，并将推特中的非结构化文本和雅虎财经中的交易数据这两个不同数据源进行特征融合。面对股票趋势预测这样一个综合性的问题，本节对该任务进行分析定义。

股票市场每天都会产生大量的交易数据，包含开盘价、最高价、最低价、收盘价、成交量等。同时，互联网的发展使得社交媒体上充斥着大量的影响股票波动的文本信息。股票预测任务是一个时间序列问题，一家股票公司在交易日 d 受到了政策福利以及历史交易的影响，其股票在未来 $[d, d+N]$ 时间间隔内受到其影响股价上涨或者下跌。本章考虑使用开盘价、收盘价作为股票价格数据特征。

给定时间序列 N、股票 s 和滞后区间 $[t-N,t-1]$，持续来看，在日期 d 内，对于某只股票 i，过去 D 日内的价格序列可以记做:

$$p_{i,d} = (p_{i,d-D}, ..., p_{i,d-1}) \tag{3.1}$$

其中，$(P_{t-N}, P_{t-N+1}, ..., P_{t-1})$ 表示股票 i 在过去第 N 个交易日的开盘价和收盘价融合后的特征。

除了股票的历史技术指标特征外，股票市场产生的相关在线内容也是影响股票趋势的重要因素。本文对股票 i 在 d 日内的推特文本视作一个文本序列，形式化表示为:
.

$$C_{i,d} = (c_{i,d-D}, ..., c_{i,d-1}) \tag{3.2}$$

其中，$(C_{t-N}, C_{t-N+1}, ..., C_{t-1})$ 表示股票 i 在过去第 N 个交易日的文本特征，可以由多条新闻消息组成。

股票趋势预测问题本质上是一个分类问题。现有的研究一般将股票下一时期的趋势根据上涨百分比分为上涨、持平和下跌三类，或者上涨下跌两个类别。对于给定日

期 t 的上涨百分比，可以通过以下公式计算：

$$RP(t) = \frac{o(t+1) - o(t)}{o(t)} \tag{3.3}$$

$O(t)$ 表示日期 t 的开盘价，$RP(t)$ 表示日期 t 的上涨百分比。

对于在第 t_1 个交易日的任一股票，当将股票趋势分为上涨下跌两类时，上涨百分比大于 0 时，即 $RP(t_1)$ 的值大于 0 时，股票趋势为上涨，y_t 的值为 1；当 $RP(t_1)$ 的值小于或等于 0 时，股票趋势为下跌或者持平，y_t 的值为 0。

$$RP(t_1) = \begin{cases} 1, RP(t_1) > 0 \\ 0, RP(t_1) <= 0 \end{cases} \tag{3.4}$$

当将股票趋势看作三类即上涨、下跌、持平时，对于在交易日 t_2 的任一股票，当 $RP(t)$ 的值大于 0 时，股票趋势为上涨，y_t 的值为 1；当 $RP(t)$ 的值小于 0 时，y_t 的值为 0；当 $RP(t)$ 的值等于 0 时，趋势为持平，y_t 的值为 2；

$$RP(t_2) = \begin{cases} 1, RP(t_1) > 0 \\ 0, RP(t_1) < 0 \\ 2, RP(t_1) <= 0 \end{cases} \tag{3.5}$$

在本章节中，股票趋势预测的目标是对日期 N 内的新闻语料库序列 [$C_{t-N}, C_{t-N+1}, ..., C_{t-1}$] 以及历史价格序列 [$P_{t-N}, P_{t-N+1}, ..., P_{t-1}$] 进行信息的有效提取，进而对两者进行信息融合，使用两种类型的来源数据作为输入，对第 t 天的股票涨跌趋势即 $RP(t)$ 的值进行预测，根据预测结果，判断股票未来趋势属于哪一类别。

$$RP(t) = f((p_{i,t-1}, ..., p_{i,t-N}), c_{i,t-1}, ..., c_{i,t-N}) \tag{3.6}$$

根据上述，本文提出了一种基于异构数据和多层注意力机制的股票预测模型 CredibleNet，将多源异构数据特征进行特征融合，然后将联合特征作为后续 LSTM 预测模型的输入，从而对股票的下一步涨跌情况进行预测。本章的剩余部分将对模型部分进行详细叙述。

3.2 模型框架

基于顺序时间序列的依赖特性，本章模型在顺序时间上下文中解释和分析股票特征，并尽可能的关注关键时间段。此外，本章的框架将更重要的新闻句与其他新闻句区分开、将重要的单词与其他单词进行区分。为了抓住这三个原则，本文设计了一个多层注意力网络，该网络在单词级、新闻级别和时间级别都加入了注意力机制。

本文提出的股票预测网络的总体架构如图3.1所示，由以下几部分组成：历史价格编码、单词级注意力层、新闻级注意力层、特征融合、时间编码器和时间级注意力层以及趋势预测网络组成。

对于股票趋势预测任务而言，新闻文本和历史价格都可以看做变化的时间序列，并且股票的波动情况与近期交易日内的新闻报道和交易信息相关性比较强。因此，为了解决影响因素间的依赖性问题，本文使用 BiLSTM 模型捕获新闻和历史价格间的长距离的依赖关系。本文考虑近 D 个交易日内的股票信息，给定时间序列的长度 D，时间序列 D 内的新闻语料库集合 N。N_i 表示在第 i 天内的所有股票相关新闻语料，N_i 包含一组长度为 L 的新闻集合。

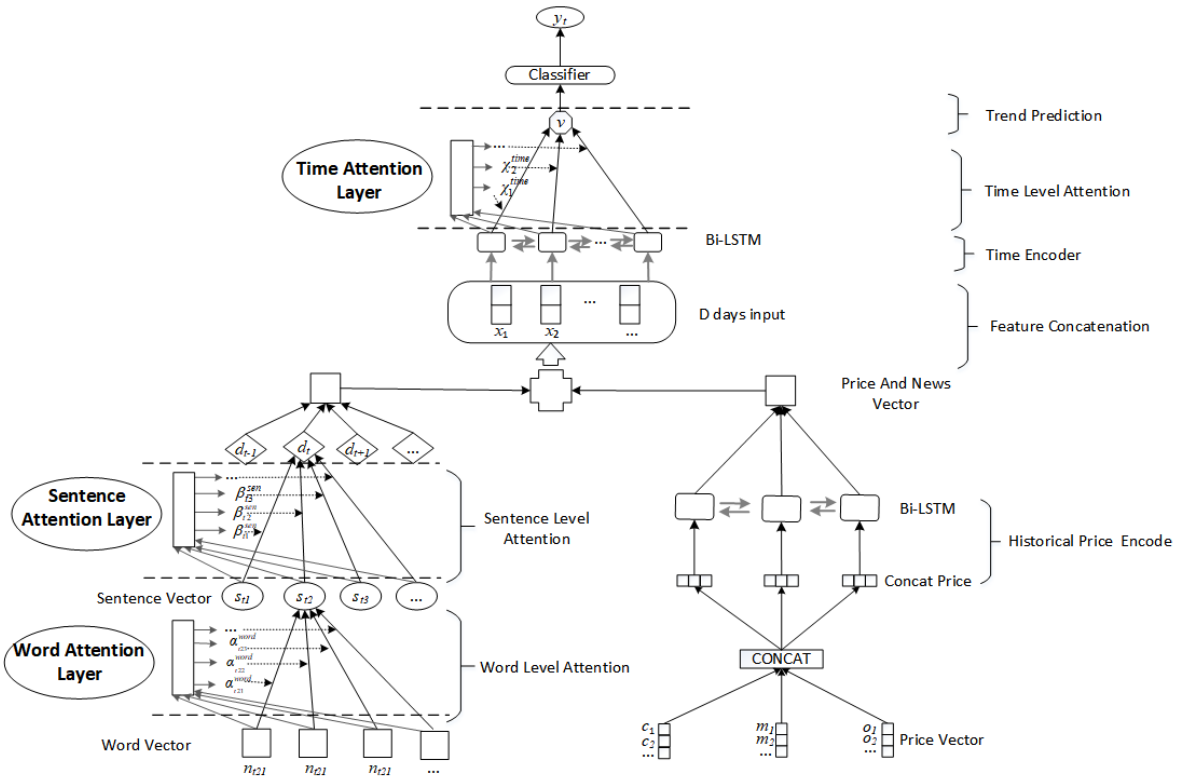

图 3.1 CredibleNet 网络整体架构

3.2.1 基于 BiLSTM 的股票历史价格编码

股票信息的时效性比较强，股票价格的波动情况与最近交易日的有关信息相关性更强，与相隔较远的交易日关联较弱。在选取历史价格信息特征方面，本章选取开盘价、收盘价以及涨跌幅作为影响股票波动的价格指标。开盘价是指每个交易日股票开盘时的价格。若开盘价没有产生，一般采用前一天的收盘价作为当天股票的开盘价。收盘价是指交易日的最后时刻的成交价格，如果当天没有买卖或者成交，则取上一个交易日的收盘价格作为当天的收盘价。涨跌幅，顾名思义就是对股票涨跌的描述涨跌幅包含收盘价、开盘价的相关信息并且描述了股票的涨跌程度，因此在模型中，本章选取的价格特征为近 N 个交易日的收盘价 $[c_{t-N}, c_{t-N+1}, ..., c_{t-1}]$、开盘价 $[o_{t-N}, o_{t-N+1}, ..., o_{t-1}]$ 以及涨跌幅度 $[m_{t-N}, m_{t-N+1}, ..., m_{t-1}]$ 作为影响股票趋势的价格影响因素，对其三者进行联合，得到融合特征 $[p_{t-N}, p_{t-N+1}, ..., p_{t-1}]$，然后将融合后的特征数据进行序列编码。根据时间序列数据的性质，使用具有双向 LSTM 单元的 BiLSTM 来递归提取特征。该体系结构解决了梯度消失和从原始数据中学习长期依赖的问题。

LSTM 结构和以及其相关变换用于通过顺序处理来分析时间序列数据，作为递归神经网络的一种变体，LSTM 使用门控机制来检查序列的状态，而无需单独的存储单元。LSTM 单元包括门和单元状态，其中单元状态由三个门控制：遗忘门，输入门和输出门。对于输入信息 p_t、h_t 和 c_t 代表输出值和单元状态，在时刻 t 时，第一步根据单元状态确定要丢弃的信息，该信息通过遗忘门的 sigmoid 单元进行处理。下一步是将新的特征信息添加到单元状态并更新原有信息，最后，确定需要输出的单元状态的特征。LSTM 的具体过程描述如下：

$$f_t = \sigma(w_f \cdot [h_{t_1}, p_t] + b_f) \tag{3.7}$$

$$i_t = \sigma(w_i \cdot [h_{t_1}, p_t] + b_i) \tag{3.8}$$

$$o_t = \sigma(w_o \cdot [h_{t_1}, p_t] + b_o) \tag{3.9}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_C * [h_{t_1}, x_t] + b_c) \tag{3.10}$$

$$h_t = o_t * \tanh(C_t) \tag{3.11}$$

本章按照前向和后向两个方向的顺序使用 LSTM 对历史数据特征序列进行编码，通过从价格数据的两个方向进行汇总来获得包含前后信息的价格特征：

$$\overleftarrow{h_t} = \overleftarrow{LSTM} p_t, t \in [1, D]) \tag{3.12}$$

LSTM 以及 BiLSTM 网络已经在许多的金融应用中取得了良好的效果。BiLSTM 通过从时间序列数据的两个方向汇总信息来获得价格可用信息，将上下文信息合并到最终的特征中。基于上述计算过程获得了价格数据的可用特征。注意力机制的出现代表了许多研究人员又向前迈出了一步。本章将单词级和句子级的注意力用于获得的股票文本信息，以获得更高效的文本特征执行预测任务。

3.2.2 基于双层注意力机制的股票文本分析

在分析利用和目标股票相关的众多新闻时，大多数研究工作对待每一条新闻都一视同仁。这种传统的方法忽略了新闻对股票的不同程度的影响，因为一些新闻显然包含重要的股票趋势信息，而有一些新闻与股票趋势的相关性较低。因此，需要增加对重要新闻的关注以提高股票预测的性能。分层注意力网络提出了一种从质量低、可行性低、质量层次不齐的新闻内容中提取可用信息的新方法，经过多层注意力机制处理后的低效新闻文本数据消除了一部分没有价值的低信息量内容。高质量、可靠的新闻输入不仅有利于提高模型运行效率，还有利于股票预测结果准确性的增加。注意力机制就是对模型的输入根据重要程度给予不同程度的权值，使得模型在进行训练预测时更加关注有价值的输入信息。一般来说，注意力机制的实现经过两个步骤：先对输入的相关信息进行注意力权重计算，再根据注意力权重计算上下文向量。在与公司股票波动相关的一系列文本语料中，每一条的新闻重要程度是不一样的，新闻级注意力层使得模型在进行股票趋势预测时会更加关注对股票波动产生关键影响的新闻信息，也就是会对重要的新闻输入给予较大的权重。同理，不同时间下的股票新闻内容对股票趋势的影响肯定也是不同的，因此，在时间级别上同样加入一个注意力层，捕获在一系列的股票运动时间中对股票趋势具有关键影响的时间点。此外，考虑到一条新闻中每个单词对句子含义贡献度也有所不同，在模型加入了一个单词级注意力层，对一条新闻中的所有单词根据其重要性赋予不同的权重。

单词级注意力: 并不是所有单词对句子含义的表示都有同样的贡献。因此，本章

引入单词级注意机制来提取对句子含义影响程度高的单词，并汇总这些信息性单词的表示形式以形成句子向量。也就是说，一条新闻中的每个单词在表达句子意思时的价值并不一样。使用词级注意力机制来赋予句子中的单词不同的权重。最后通过加权求和的方式将所有单词聚合成句子向量 s_{ti}^{sen}, 具体为：

$$u_{tim}^{word} = sigmoid(W_w n_{tim} + b_w) \tag{3.13}$$

$$\alpha_{tim}^{word} = \frac{exp(u_{tim}^{word})}{\sum_k u_{tik}^{word}} \tag{3.14}$$

$$s_{ti}^{sen} = \sum_k \alpha_{tim}^{word} n_{tim} \tag{3.15}$$

将嵌入层得到的词向量 n_{tim} 通过单层感知机获得单词级隐层表示 u_{tim}^{word}，通过 softmax 函数得到归一化的注意力权重 α_{tim}^{word}。通过计算单词向量的加权和作为新闻的句子向量 s_{ti}^{sen}, 表示为在时间 t 内，第 i 条新闻中每个单词向量的加权和。

新闻级注意力: 并非所有句子都对预测股票趋势做出了同等贡献，因此本章引入了一种句子级注意力机制来汇总依据关注值加权的股票文本，以重点关注提供关键信息的股票文本信息。具体来说，为了提取到对股票预测产生有效影响的高价值新闻信息，本章使用了句子级注意力机制并引入了句子级别的上下文向量 ν_{tj}^{sen}，使用该向量来衡量一天内不同新闻句子的重要程度，就产生了：

$$\nu_{tj}^{sen} = sigmoid(W_h s_{tj}^{sen} + b_n) \tag{3.16}$$

$$\beta_{tj}^{sen} = \frac{exp(\nu_{tj}^{sen})}{\sum_b \nu_{tb}^{sen}} \tag{3.17}$$

$$d_t^{sen} = \sum_j \beta_{tj}^{sen} s_{tj}^{sen} \tag{3.18}$$

同理，ν_{tj}^{sen} 代表了新闻层的注意力，β_{tj}^{sen} 是经过标准化以后的注意力权重，向量 d_t^{sen} 是多个新闻向量的加权和，代表着第 t 天内的所有股票相关新闻信息。d_t^{sen} 经过两层注意力机制，即词级注意力和句级注意力，两次加权求和得到的信息。因此，可以得到语料库向量 $D = [d_t^{sen}]$，$t \in [1, N]$ 的时间序列。注意力层可以为可靠且信息丰富的股票文本分配更多的注意力。

3.2.3 异源数据特征融合

在深度神经网络学习的过程中，不同数据的特征表达起着关键作用。特征融合可以看作是多个方面的结合，能够获得更加有效的表示，它可以理解为将各种因素结合起来提升预测性能。股票的涨跌趋势受多方面因素的影响。由于在股票领域特征之间是相互影响和相互联系的，而且不同的特征对股市的影响也是不同的。股票各种相关信息的有效整合，可以改善新闻表达，进而提高股票预测性能。股票研究者们发现新闻文本特征与股票价格特征结合能提高股票预测的精度。在预测股票市场时，输入特征可以考虑使用历史交易数据和新闻文本数据的结合特征进行趋势预测，这将有利于提高股票预测结果的精度。为了构建不同特征的前后联系以及不同特征的重要性，本文提出了基于 BiLSTM 和双层注意力机制的的融合方式。在本文的方法中，首先采用 BiLSTM 来构建不同历史价格特征的前后关系，然后采用双层注意力机制对低质量的股票相关评价文本赋予不同的权重，最后对两种特征进行融合。本文中的数据融合指的是实现融合股票相关异构数据源，为股票模型输入提供更准确、更高效的股票特征。

给定过去 N 个交易日的收盘价、开盘价以及涨跌幅度经过 BiLSTM 捕捉时间依赖性后得到的历史价格量化特征 [p_{t-N},p_{t-N+1},...,p_{t-1}]。同时，股票相关的句子信息经过单次级注意力机制和句子级注意力机制选取关键特征后得到股票文本量化特征 [d_{t-N}^{sen},...,d_{t-1}^{sen}]。对于以上特征序列, 将量化后的股票文本特征序列 [d_{t-N}^{sen},...,d_{t-1}^{sen}] 和历史价格特征序列 [p_{t-N},p_{t-N+1},...,p_{t-1}] 拼接在一起得到一个新的向量 dp。

时间编码: 对于股票趋势预测来说，股票价格和股票文本信息都是时间序列数据，长短时记忆网络可以很好地描述随时间变化的时间序列。因此，对融合后的向量 dp 进行时间序列编码，捕获序列依赖，使用 BiLSTM 对向量进行编码。依据上述提到的单词序列以及文本序列编码过程，可以得到时间序列编码过程：

$$\overrightarrow{h}_t^{concat} = \overrightarrow{LSTM}(dp), t \in [1, D] \tag{3.19}$$

$$\overleftarrow{h}_t^{concat} = \overleftarrow{LSTM}(dp), t \in [1, D] \tag{3.20}$$

$$h_t^{concat} = [\overrightarrow{h}_t^{concat}, \overleftarrow{h}_t^{concat}] \tag{3.21}$$

3.2.4 时间级注意力处理融合数据

考虑到不同时间发布的新闻和历史价格对股票波动做出的贡献并不一样。时间层的注意力机制可以使模型在一个长周期内，提取出影响股票趋势的重要时间内所包含的重要信息，有利于区分时间差异。如下所示：

$$x_s^{time} = sigmoid(W_h h_t^{concat} + b_h) \tag{3.22}$$

$$\chi_s^{time} = \frac{exp(\theta_s x_s^{time})}{\sum_c exp(\theta_c x_c^{time})} \tag{3.23}$$

$$V^{time} = \sum_s \beta_s h_t^{concat} \tag{3.24}$$

其中 θ_s 参数在 softmax 层中起的作用是选择哪个日期更加重要，x_s^{time} 用于编码语料库向量的潜在表示。通过 softmax 层将两者结合起来得到注意力向量 χ_s^{time} 来对不同的时间进行区分。然后利用注意力向量计算加权和得到 V^{time}。

3.2.5 趋势预测

本章将经过处理后的融合特征向量 V^{time} 作为输入，该向量结合了股票文本信息和股票历史价格的相关信息，然后使用图3.2所示的多层 LSTM 神经网络作为分类器将其用于股票分类，输出下一步股票趋势的分类可能情况 y_{class}，以二分类为例，y_{class}=0 表示下一阶段的股票预测结果为下跌，y_{class}=1 表示股票趋势很大可能会为上涨。在时刻 t 第 n 层神经网络的循环神经元状态 s_t^n 具体计算公式如下：

$$tanh = \frac{e^x - e^{-x}}{e^x + e^(-x)} \tag{3.25}$$

$$\begin{bmatrix} i_n^t \\ f_n^t \\ 0_n^t \\ \overline{s}_n^t \end{bmatrix} = \begin{bmatrix} \theta \\ \theta \\ \theta \\ tanh \end{bmatrix} \begin{bmatrix} W_{i,x}^n W_{i,r}^n \\ W_{f,x}^n W_{f,r}^n \\ W_{o,x}^n W_{o,r}^n \\ W_{\overline{s},x}^n W_{\overline{s},r}^n \end{bmatrix} \begin{bmatrix} r_t^{n-1} \\ r_{t-1}^n \end{bmatrix} \tag{3.26}$$

$$y_{class} = RELU(W * r_t^n + b) \tag{3.27}$$

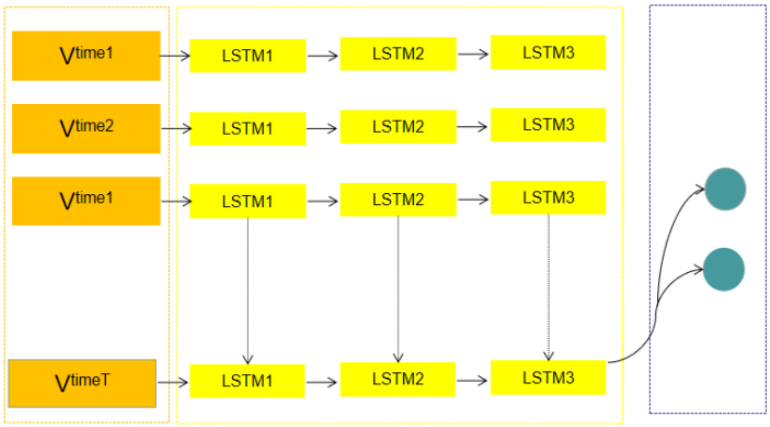

图 3.2 多层 LSTM 神经网络图

3.3 实验结果与分析

3.3.1 数据集描述

为了更方便地将本章方法与现有方法进行比较，选择使用在[41] 中的数据集上进行股票的趋势预测，该数据集可以被视为股票走势预测任务中的代表性参考，一些研究者在此数据集上进行了一些早期工作和新进展的实验[41, 61–63]。此外，根据其划分原则，将股票分为 9 个行业：基础材料、消费品、医疗保健、服务、公用事业、企业集团、金融、工业产品和科技。考虑到高成交量的股票更容易引起人们的讨论，该数据集包括了 8 只企业集团的股票和其他 8 个行业的资本规模排名前 10 的股票。除了推特评论，这 88 只股票的价格数据也被额外纳入最终数据集。时间范围为 2014 年 1 月 1 日至 2016 年 1 月 1 日。每支股票都有相对应的两个重要信息，包括来自推特的推文数据以及来自雅虎财经的价格数据，该数据集中的推文已经进行了预处理。为了更好的体现本文模型的稳健性，在实验部分将会对股票趋势的二分类和三分类均进行研究。

当预测目标是对股票趋势变化进行三分类时，为了给定分类问题的标签，根据移动百分比设置了两个特定的阈值-0.5% 和 0.55% 来进行类别划分。移动百分比小于等于-0.5% 和移动百分比大于等于 0.55% 的样品分别用 0 和 1 标记，而两个阈值之间的即为持平，用标签 2 进行标记。选择的两个阈值应使得三个类别大体平衡，在整个数据集中有 23282 个预测目标，上涨、持平、下跌三个个类别分别占 33.04%、32.58% 和 34.38% 比例大致相同。接着对数据集在时间范围上根据以下原则进行拆分，2014 年 1 月 1 日到 2015 年 8 月 1 日之间 78.6% 的股票数据用于训练。2015 年 8 月 1 日到

2015 年 10 月 1 日之间 8.5% 的股票数据用于做验证集。2015 年 10 月 1 日到 2016 年 1 月 1 日期间的 12.9% 股票数据用于测试。按照以上原则进行划分，三个类别的数据比例大致相等。

当将股票趋势预测问题看作二分类问题时，预测类别为上涨和下跌。上涨趋势的股票标签设置为 1，下跌的股票标签设置为 0。与上述提到的三分类相比，删除了移动百分比在两个阈值之间的 32.68% 的数据。划分数据集时，将 2014 年 1 月 1 日到 2015 年 8 月 1 日之间 76.2% 的股票数据用于训练。2015 年 8 月 1 日到 2015 年 10 月 1 日之间 9.7% 的股票数据用于做验证集。2015 年 10 月 1 日到 2016 年 1 月 1 日期间的 14.1% 股票数据用于测试。按照以上原则进行划分，上涨和下跌两个类别的数据比例大致相等。

3.3.2 实验设置

使用过去 5 天的价格和文本信息来预测下一天的涨跌情况，在一个批次中使用 32 个混合样本，将最大日期长度设置为 40，最大新闻长度设置为 30，并删除超过限制的样本。由于批处理样本中的所有文本数据和历史交易数据都同时输入到模型中，考虑到内存成本，将单词嵌入大小设置为 50。本章模型在 CPU 上运行。模型中权重矩阵默认使用扇入技巧初始化，偏差用零初始化。使用 Adam 优化器对模型进行训练，初始学习率为 0.001，使用 0.3 的输入退出率 (input dropout) 来调整潜在变量。

3.3.3 评价指标

在股票趋势的分类问题中，类别划分比例大致一样，本章节采用以下在股票预测领域常用的预测评价指标[64] 来对本文中的模型性能进行评价。下述的衡量标准都是基于混淆矩阵表3.1来对分类模型进行评价的，其中 tp 代表分类中的真正性的样本数，tn 表示真负性的样本数，fp 表示假正性的样本数，fn 表示假负性的样本数。

表 3.1 混淆矩阵

实际结果	预测结果	
	正例	反例
正例	真正例 (tp)	假反例 (fn)
反例	假正例 (fp)	真反例 (tn)

准确率 (accuracy)：准确率是衡量分类模型性能好坏最直观的评价指标，它是正

确预测的数据量与所有参与预测的数据量的比值。

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{3.28}$$

精确率 (precision)：精确率就是在预测结果为正例的样本数据中，真实数据也为正例所占的比重。

$$precision = \frac{tp}{tp + fp} \tag{3.29}$$

召回率 (recall)：召回率表示在真实数据为正例的所有样本中，预测结果也为正例的样本所占的比例。

$$recall = \frac{tp}{tp + fn} \tag{3.30}$$

F1 分数 (F1-Score)：F1-Score 是精确率和召回率的加权平均值。

$$F1 - Score = 2\frac{precision * recall}{precision + recall} \tag{3.31}$$

马修斯相关系数 (MCC)：MCC 作为评价分类问题分类性能的指标，通常可以有效避免由于数据不平衡而产生的偏差。取值范围为 [-1,1]，取值为 1 时为准确预测，取值为 0 代表预测结果不如随机预测，取值为-1 表明预测结果与实际结果完全相反。

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{3.32}$$

以上的点预测评价指标只能用于衡量模型分类的预测能力，但是当模型应用在在真实的金融市场时，无论点预测模型多么优异，总会存在不确定性导致预测结果出现偏差。因此还应对模型的可靠能力给出一个评价。不确定性描述了模型预测结果的可靠程度，如果不确定性值趋近于 1，表明预测结果完全不可靠，相反，如果不确定值靠近 0，则说明预测结果相当可靠。当不确定性非常高时，即使上述点预测的评价指标表现优异，则投资者在做投资决策时应更加慎重。

使用均值和偏度来对模型的不确定性的性能进行评价分析。较小的均值表明不确定性较小，表明模型性能越好。偏度是一种信息统计指标，偏度的绝对值越小，性能越好, 定义如下：

$$SK = \frac{1}{n}\sum_{i=1}^{n}[(\frac{X_i - \mu}{\sigma})^3] \tag{3.33}$$

μ 表示均值，σ 表示方差。

3.3.4 消融实验预测结果与分析

为了对 CredibleNet 中各个组成部分进行分析，本文对以下五个 CredibleNet 变体模型进行对比实验，分析了各个部分对股票趋势预测的有效程度。

PriceAnalyst(PAna)：为了证明特征融合的有效性，在此变体模型中仅使用经过双向编码捕捉依赖性以及时间注意力机制处理的股票历史价格作为输入进行预测。

NewsAnalyst(NAna)：在此模型中，使用只经过两层注意力机制的处理的股票的相关文本数据作为输入进行预测，并不使用融合特征。

WordAttAnalyst(WAtt)：为了评估单词注意力层，在该模型中仅加入单词级的注意力机制进行预测，而不考虑其它注意力层。

NewsAttAnalyst(NAtt)：为了评估新闻注意力层，通过添加单个新闻层次的注意力层来处理语料库向量。

TimesAttAnalyst(TAtt)：为了评估时间级注意力层，仅添加时间注意力层进行预测。

CredibleNet(Cre)：本章提出的包含有三层注意力机制以及多次编码过程的股票预测网络。

模型性能情况如表3.2、表3.3所示。为了更加充分的体现所提出模型的适用性，本章节对股票预测的二分类以及三分类情况均进行了研究。通过表中变体模型在股票趋势三分类和二分类中的评价指标对比情况可以看出，仅仅使用价格或者文本对股票进行预测，效果并不理想，而且相比较使用推特新闻数据作为数据集，使用历史价格进行股票预测效果会更好，这是由于推特新闻中存在大量的低质量、可用性差的无效数据所造成的，这体现了对推特新闻进行处理的重要性。将价格信息与文本信息进行融合以后，再加入单词级、句子级、时间级任意一层的注意力机制，效果便得到了提升。当加入三层注意力机制后，可以看出模型性能得到进一步的提升，体现了考虑多维度因素以及多层注意力机制进行股票预测的优势。

性能讨论：根据 priceanalyst 和 newsanalyst 模型的结果，可以看出仅仅使用历史价

表 3.2 变体模型性能情况 (三分类)

Methods	price	news	word attention	sentence attention	time attention	Metrics				
						accuracy	precision	recall	F1-score	MCC
PriceAnalyst	✓	-	-	-	-	38.497	0.67	0.38	0.24	0.120
NewsAnalyst	-	✓	-	-	-	36.104	0.13	0.36	0.19	0.000
WordAttAnalyst	✓	✓	✓	-	-	39.993	0.58	0.40	0.28	0.135
NewsAttAnalyst	✓	✓	-	✓	-	39.062	0.62	0.39	0.32	0.146
TimeAttAnalyst	✓	✓	-	-	✓	40.559	0.50	0.41	0.33	0.138
CredibleNet	✓	✓	✓	✓	✓	**54.654**	**0.8**	**0.55**	**0.54**	**0.450**

表 3.3 变体模型性能情况 (二分类)

Methods	price	news	word attention	sentence attention	time attention	Metrics				
						accuracy	precision	recall	F1-score	MCC
PAna(Ours)	✓	-	-	-	-	52.826	0.74	0.53	0.42	0.215
NAna(Ours)	-	✓	-	-	-	52.394	0.27	0.52	0.36	0.000
WAtt(Ours)	✓	✓	✓	-	-	56.383	0.72	0.56	0.49	0.265
NAtt(Ours)	✓	✓	-	✓	-	54.255	0.62	0.54	0.48	0.163
TAtt(Ours)	✓	✓	-	-	✓	54.222	0.56	0.54	0.52	0.113
Cre(Ours)	✓	✓	✓	✓	✓	**60.372**	**0.74**	**0.60**	**0.55**	**0.337**

格交易数据或者新闻数据作为输入并不能够产生令人满意的预测结果。与 priceanalyst 和 newsanalyst 相比，WordAttAnalyst、NewsAttAnalyst 和 TimeAttAnalyst 使用两种市场信息的融合特征作为输入，分别区分了不同单词、新闻、日期的不同影响，在预测任务中取得了更好的效果。与上述模型相比，CredibleNet 由于考虑到特征融合，并引入了三层的注意力机制和 BiLSTM 编码方式对市场信息进行处理，获得了更好的性能表现。

3.3.5 基线模型预测结果与分析

为了证明本文提出模型的有效性，将所提出模型与以下现有的股票预测方法的性能进行对比：

多层感知机[65](MLP)：使用多层感知机作为分类器，使用特征融合后的数据作为输入源。

长短时记忆网络[66](LSTM)：将特征融合后的数据作为输入，使用两层的 LSTM 模型进行分类预测。

双向长短时记忆网络[67](BiLSTM)：为了评估双向设置的有效性，使用 BiLSTM 对股票进行预测，输入数据与 MLP、LSTM 相同。

多层注意力网络[39](HAN)：一个具有新闻级和时间级层次注意力机制的深度神经网络，以文本数据为输入通过计算当前全局张量中的相应嵌入的加权和来学习更高级别的表示。

基线模型在二分类与三分类情况下的性能分别如下表3.4以及表3.5所示。

表 3.4 基线模型性能情况 (三分类)

Variations	accuracy	precision	recall	F1-score	MCC
MLP	45.911	0.44	0.46	0.39	0.210
LSTM	48.105	0.49	0.48	0.45	0.233
Bi-LSTM	51.031	0.63	0.51	0.44	0.311
HAN(Hu et al.,2018)	52.294	0.77	0.52	0.44	0.368
CredibleNet	**54.654**	**0.8**	**0.55**	**0.54**	**0.450**

在股票预测三分类的情况下，所有基线方法的性能比较结果如表3.4所示，为了更好的可视化结果，本章将结果展现在并列柱状图中。通过图3.3可以更加直观的观察到指标的对比情况，图3.3展现了在各项评价指标中不同基线模型的结果对比。根据表3.4可以看出在基线模型中的最优指标相比，本章模型在准确性、精确性、召回率、F1-分数、MCC 指标方面相比于其他模型分别提高了 2.36、0.03、0.03、0.09、0.079。

表3.5和图3.4展示了在股票预测二分类时基线模型的表现情况，根据实验结果可以看出在基线模型中，HAN 模型与传统的神经网络模型相比，具有更好的性能表现，与 HAN 模型相比，CredibleNet 模型在准确性、精确性、召回率、MCC 方面分别提高了 2.028、0.03、0.02、0.054。

表 3.5 基线模型性能情况 (二分类)

Variations	accuracy	precision	recall	F1-score	MCC
MLP	52.560	0.52	0.53	0.52	0.046
LSTM	52.926	0.53	0.53	0.53	0.059
Bi-LSTM	55.785	0.58	0.56	0.55	0.138
HAN(Hu et al.,2018)	58.344	0.71	0.58	0.53	0.283
CredibleNet	**60.372**	**0.74**	**0.60**	**0.55**	**0.337**

根据上述 CredibleNet 模型在股票预测二分类以及三分类中与其他基线模型的性能比较可以看出，CredibleNet 模型在所有评价指标中明显优于其他基线方法。同时，可以看出，相比于传统的神经网络模型，加入注意力机制后的网络的预测效果明显提高。这表明了使用注意力机制用于股票预测能够显著的提高准确性等各项评价指标。对比结果表明了本文的预测模型 CredibleNet 在预测趋势时的表现效果良好。

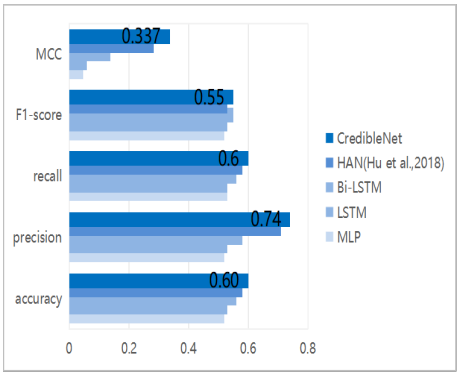

图 3.3 基线模型的指标对比情况 (三分类)

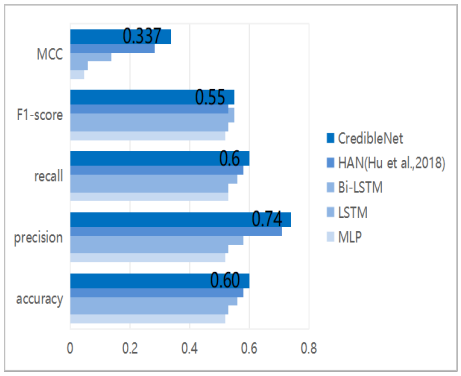

图 3.4 基线模型的指标对比情况 (二分类)

性能讨论：在神经网络模型中，MLP 的性能显然比其他的模型要差，可能是因为这种方法的输入没有按照顺序上下文进行组织，这个结果在一定程度上证明了使用 LSTM 的重要性，LSTM 可以按照上下文顺序处理信息。通过 LSTM 与 BiLSTM 的结果对比可以看出，BiLSTM 优于 LSTM，这表明了双向编码的优越性，其可以利用过去和将来的信息进行预测，有利于性能的提升。HAN 模型相比于传统的神经网络模型，加入了新闻级和时间级的两层注意力机制，可以看出，相比于没有引入注意力机制基线模型，HAN 模型的性能得到了明显提升，说明了区分不同新闻、时间影响的有效性。CredibleNet 模型可以实现比以上所有模型更好的性能，表明了本文所提出模型的有效性。

3.3.6 不确定性分析

目前，很多股票预测的研究都是基于点预测指标，本文在点预测之后增加了不确定性分析部分，采用均值和偏度两个指标来评价不确定性，本节讨论不确定性的影响。

在表3.6中对本章的基线模型和变体模型进行股票预测三分类时的不确定进行评价，根据表中的均值以及偏度结果可以看出，在经典的神经网络模型 MLP、LSTM、BiLSTM 中，LSTM 是均值最低且偏度绝对值最小的基线模型，说明 LSTM 模型的预测可靠性比较好，因此考虑在 CredibleNet 中使用 LSTM 作为股票趋势的分类器。

表 3.6 不确定性评估 (三分类)

	MLP	LSTM	Bi-LSTM	HAN	PAna	NAna	WAtt	NAtt	TAtt	Cre
Mean	0.870	0.714	0.749	0.511	0.809	0.872	0.633	0.836	0.668	**0.470**
Skewness	-1.683	-0.926	-0.987	0.507	-1.879	-1.712	-0.575	-0.644	-0.415	**0.142**

根据图3.5可以看出，在基线模型中 MLP 模型在均值和方差方面表现均最差，在变体模型中 NAtt 即仅使用推特新闻作为输入进行股票预测的模型表现最差，推测这是由于推特新闻本身就具有较大的不确定性所造成的。HAN 模型的均值和偏度性能表现均明显好于传统的神经网络模型，这可能 HAN 模型以加入了两层注意力机制造成的，而 CredibleNet 模型在均值和偏度方面均达到了最佳性能，分别以-0.244、-0.041 的均值和-0.784、-0.365 的偏度值优于 LSTM 和 HAN。

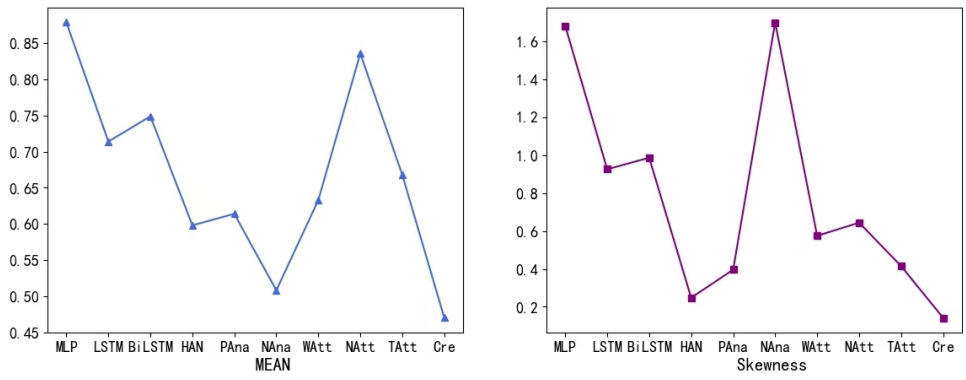

图 3.5 不同预测模型的不确定性统计 (三分类)

根据表3.6中变体模型的不确定性结果，可以看出在变体模型中，当只使用历史交易数据或者只使用股票推特数据进行预测时，均值和偏度表现一般，三个注意力机制

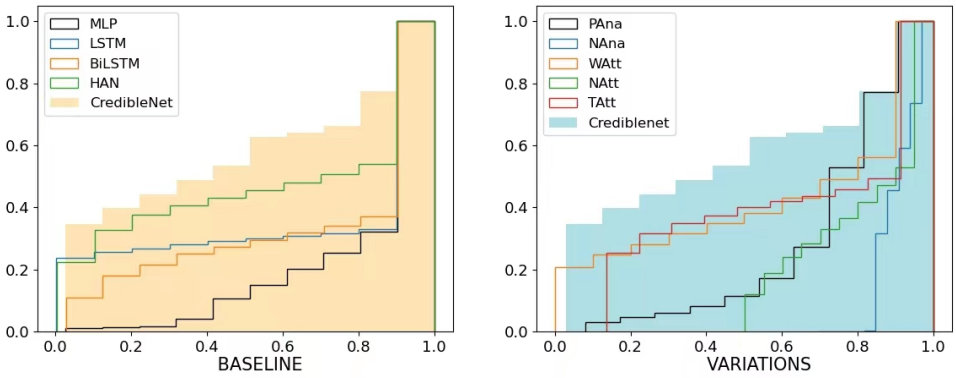

图 3.6 基线模型与变体模型的不确定性 CDF 图 (三分类）

的模型与一个注意力机制模型的对比可以看出，三层注意力模型的可靠性得到明显提升。此外根据图3.6展现的基线模型和变体模型的不确定一致性累计分布函数图 (CDF 图)，可以看出不确定性数值的累积分布函数，可以看出，CredibleNet 模型在不确定性方面表现出了强大的优势。与其他常见的股票预测模型相比 CredibleNet 大部分的不确定值都小于 0.8，与其他的模型相比不确定性更低，说明所提出的模型具有良好可信度。

表 3.7 不确定性评估 (二分类)

	MLP	LSTM	Bi-LSTM	HAN	PAna	NAna	WAtt	NAtt	TAtt	Cre
Mean	0.959	0.601	0.614	0.750	0.809	0.971	0.834	0.718	0.650	**0.598**
Skewness	-4.351	-0.362	-0.422	-0.958	-1.879	-2.108	-1.393	-0.456	-0.436	**-0.114**

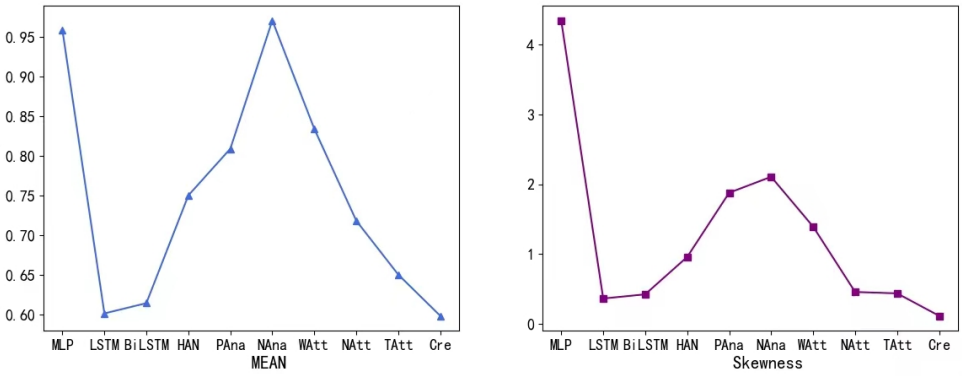

图 3.7 不同预测模型的不确定性统计 (二分类）

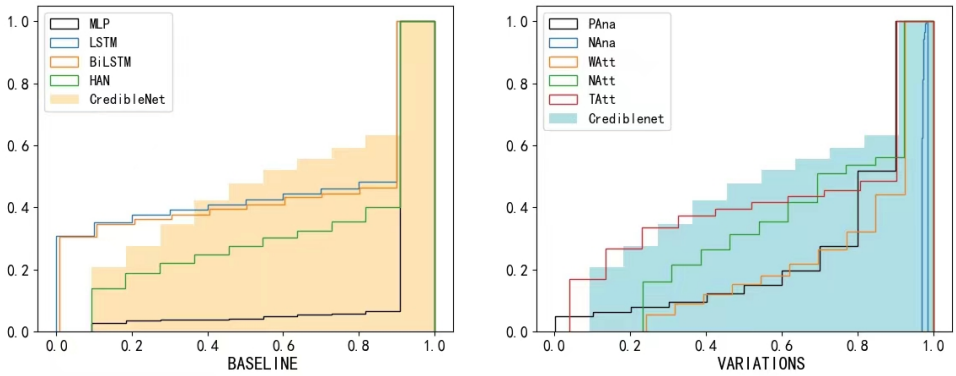

图 3.8 基线模型与变体模型的不确定性 CDF 图 (二分类）

在表3.7、图3.7、图3.8中基线模型和变体模型在股票二分类时的不确定评价可以看出，与股票预测三分类的结果类似，LSTM 仍然是神经网络模型中均值最低且偏度绝对值最小的基线模型，印证了 LSTM 模型良好的预测可靠性，MLP 模型也仍然是基线模型中均值和方差表现最差的网络模型，在变体模型中 NAtt 模型表现最差。同三分类中的表现一样，HAN 模型的均值和偏度性能表现明显好于传统的神经网络模型，而 CredibleNet 模型在均值和偏度方面同样达到了最佳性能。此外根据图3.8展现的二分类的 CDF 图可以看出模型 CredibleNet 在二分类中大部分的不确定值都小于 0.6，与其他的模型相比具有更好的可信度。

3.4 本章小结

为了更好的利用新闻文本和交易信息，提高多种数据源的可用性，本章提出了一种基于异构数据和多层注意力机制的预测方法，考虑到时间序列因素，本文使用 BiLSTM 模型捕捉历史交易信息中的长期依赖关系，获取高效的量化特征。此外，为了从新闻文本中获取高质量、可用的向量特征，采用了三层注意力机制与 BiLSTM 网络相结合的模型。在 3.1 小节中，简述了股票预测的任务描述。在 3.2 小节中，详细介绍了基于注意力机制的股票趋势预测网络架构，包括历史价格编码、单词级注意力、新闻级注意力、特征融合、时间编码、时间级别注意力以及趋势预测七个部分。

在历史价格编码部分，对于历史价格数据中的特征获得，利用 BiLSTM 网络对价格数据的前后信息进行汇总，得到量化特征。对于新闻文本向量化，由于股票相关文本的可用性低、质量参差不齐，本文采用两层注意力的机制，获取单词级别和句子级

别的新闻的向量表示，在单词级和新闻级注意力部分对此进行详细描述。在时间级注意力部分，考虑到一天中存在多条新闻消息，通过时间级注意力机制对融合后的历史交易数据和股票文本数据进行处理，获取天级新闻向量表示。最后，将上述部分进行组合，建立趋势预测模型。经过上述模型，本章节完成了对多源数据的融合以及股票数据的高质量特征提取两大挑战，有效的提高了股票趋势预测的准确性。实验表明，对比分析各消融方法以及不同基线模型的实验结果，CredibleNet 能够有效利用样本语义信息提高股票趋势预测的精度。

第 4 章　基于特征选择和不确定性量化的股票价格预测模型

在现有的股票趋势预测中，大多模型都严重依赖于股票指标的选择，选择指标时非常依赖于专业金融知识，很难进行迭代优化。因此本文首先使用极度随机树模型对股票的历史价格特征进行选择。

虽然利用深度神经网络对股票进行预测的研究已经取得较好的成果，但大多数深度神经网络对股票市场趋势的突然波动不够敏感。针对上述问题，本文提出了一种基于特征选择和不确定性量化的股票预测网络，通过利用混合高斯模型对神经网络输出的 logit 进行建模拟合，然后根据高斯混合模型的概率密度函数得到数据的不确定性分数，根据不确定性量化分数决定是否将该数据输入后续模型进行训练预测。相比于单纯利用深度学习方法进行股票的趋势预测，基于不确定性量化的方法通过科学合理的度量发现目前训练集以及测试集中异常的数据特征，从而提高预测模型的性能。

此外，现有的大部分模型在选取股票趋势预测的决策模型时，大多采用传统的深度学习网络进行预测，本文使用 SDENet 模型作为股票趋势预测模型，SDENet 模型从动力学角度出发，将深度神经网络的正向传播看作动态系统的状态转换，由获取高预测精度的 drift net 网络和挖掘不确定性的 diffusion net 网络构成。SDENet 能够从大量具有高噪声的数据中选取低不确定性的可靠数据进行训练，进而获得良好的股票预测效果。

4.1　极端随机树选择价格特征

为了提高股票预测模型的预测准确率并降低误差，本节使用极端随机树的特征筛选算法对股票的历史交易价格数据进行选择。极端随机树是把大量的决策树模型集成在一起后得到的集成学习算法。相比于传统决策树选择算法，极端随机树的极度随机性使得模型的方差相对小于传统决策树模型。极端决策树的关键在于如何选择分裂节点：从全体样本中随机选择 k 个特征，每个特征随机选择一个分裂节点，从而得到 k 个分类节点，然后计算这 k 个分裂节点的重要性，然后选择得分最高的作为分裂节点。

极端随机树先对输入的特征值进行预先处理，然后计算数据中各个特征的信息增益率，最后依据信息增益率对所有特征进行排序，并将低相关性的特征进行保留。以股票价格特征中的开盘价 O 为例，数据集 D 中 O 的信息增益率 $Gainratio(D,O)$ 计

算公式为：

$$Gainratio(D,O) = \frac{Gain(D,O)}{IV(O)} \tag{4.1}$$

其中，信息熵 $Gain(D,O)$ 为：

$$\begin{aligned} Gain(D,O) &= Ent(D) - Ent(D|O) \\ &= -\sum_{d\in D} p(d)logp(d) - \sum_{o\in O} p(o) * Ent(D|O=o) \\ &= -\sum_{d\in D} p(d)logp(d) - \sum_{o\in O} p(o) * [-\sum_{d\in D} \\ & p(d|O=o)logp(d|O=o)] \end{aligned} \tag{4.2}$$

其中，$Ent(D)$ 为划分前的熵，$Ent(D|O)$ 为划分后的熵。$p(o)$ 为特征开盘开盘价各类划分的频率。

常见的股票历史交易价格数据特征除了开盘价外还包括涨跌幅、收盘价、最高价、最低价以及交易量等。本文使用极端随机树算法对上述股票价格数据特征进行处理，依次计算各个属性的信息增益率，从而得到各种特征影响股票未来趋势的特征重要性分数。

4.2 基于 logit 不确定性量化处理股票数据

在进行股票预测任务时，关键在于如何判别数据集中的噪声数据和异常值数据。通过对股票数据进行不确定性量化，得到不确定性分数。在模型预测股票未来趋势时，对于具有高不确定性的股票相关特征应该尽量舍弃，关注更加确定的特征向量。将不确定性量化的思想引入到股票预测模型中，有利于通过增加股票数据的可用性进而提升模型的准确性[68]。

目前常见的许多评估分类不确定性的方法都依赖于从神经网络生成的 softmax 概率。然而，这些概率或这些概率的熵是非常不可靠的。在本文中，基于神经网络的 logit 输出得出了一种可靠的度量数据不确定性的方法，该度量可以帮助检测数据集中的数据是否可靠。基于 logit 的不确定性量化方法适用于产生 logit 的任何模型，通俗来说，logit 输出捕获数据不确定性，如果 I 类与 II 类相似，但与 III 类不同，则 I 类 logit 输出的 II 类 logit 值高于同一 logit 输出的 III 类 logit 值。该不确定量化方法的关键思想是使用高斯混合模型来对每个类别正确预测的训练样本的 logit 输出进行建模，

并基于高斯混合模型的概率密度函数对不确定性值进行建模。

利用高斯混合模型进行建模：来自一个类的样本应该是相似的，logit 值可以捕获不确定性，这意味着同一类的 logit 输出应在 logit 向量的每个维度中共享相似的 logit 值。因此，如果样本的 logit 值与预测类的已知 logit 值不同，那么估计预测的可信度较低，也就是不确定性值较高。因此，使用正确预测的训练数据的原始 logit 输出。使用高斯混合模型对分类正确的训练数据的 logit 进行建模。贝叶斯准则作为一种确定数据集最优成分数量的方法，常被用于确定高斯混合分布的模型数量。遍历每个类别依据肘部准则来确定合理的高斯混合分布组件数量。

$$BIC = ln(m)k - 2ln(L) \tag{4.3}$$

上述公式中，m 表示样本的数量，k 表示模型参数个数，L 代表基于样本拟合模型的最大似然函数。根据公式 (4.3) 可以看出，BIC 是由似然函数和一项惩罚项组成。惩罚项与模型拟合的参数选择有关。BIC 最终选择的参数会使得模型边际似然函数最大。根据肘部准则得到组件数量，确定高斯混合模型。数据集训练了一个分类模型，对于类别 i，选择此类别中正确预测的训练数据，并使用高斯混合模型对这些数据输出的 logit 进行建模。进而得到高斯混合分布的最大概率密度函数，为计算模型的不确定性提供准备。

不确定性得分：接下来的想法是将不确定性度量基于高斯混合模型的概率密度函数：密度函数值越大，不确定性越小。为了能够在数据集中观察到的密度函数的极小值，本文设计了一个得分函数以使不确定性值更易于使用。最后，使用 sigmoid 函数将从得分函数获得的值映射到 0 到 1 之间的范围。

假设高斯混合模型具有概率密度函数 pdf_i，定义特征向量 X 的得分 $s_i(x)$ 为：

$$s_i(x) = ln(max(pdf_i(t))) - ln(pdf_i(x)) \tag{4.4}$$

根据公式 (4.4) 可以看出，如果 X 是一个随机向量那么 $s_i(X)$ 也是一个随机向量，将得分 $s_i(X)$ 的第 q 个百分位数记为 s_{iq}。使用 sigmoid 函数将 $s_i(x)$ 映射到 [0,1] 之间：

$$g_i(s) = \frac{1}{1 + e^{-c_{i1}(s - c_{i2})}} \tag{4.5}$$

c_{i2} 和 c_{i1} 使用四个超参数 $0 \leq u_1 \leq 1$，$0 \leq u_2 \leq 1$，$0 \leq q_1 \leq 1$，$0 \leq q_2 \leq 1$ 确定，将 s_i 的 q_1 分位数映射到 u_1，q_2 分位数映射到 u_2，即 $g_i(s_{iq1}) = u_1$，$g_i(s_{iq2}) = u_2$。

$$f(x) = \begin{cases} c_{i2} = \dfrac{s_{iq2} ln(u_1^{-1} - 1) - s_{iq1} ln(u_2^{-1} - 1)}{ln(u_1^{-1} - 1) - ln(u_2^{-1} - 1)} \\ c_{i1} = \dfrac{-ln(u_2^{-1} - 1)}{s_{iq2} - c_{i2}} \end{cases} \tag{4.6}$$

当遇到一个被归类为 i 类的新数据样本 x 时，其不确定值为：$u(x) = g_i(s_i(x))$。对于预测为 i 类的 x_1 和 x_2，如果 $g_i(x_1) > g_i(x_2)$，根据单调性可以得知，$u(x_1) > u(x_2)$。

利用不确定性得分提高模型预测准确度：基于上述描述，本章对分类模型的输入数据进行不确定性量化，对股票相关数据给予一定的不确定性分数，低分数据意味着数据可靠性较高，采用低不确定性分数的数据参与股票趋势预测，进而提高股票预测的准确性。具体流程如算法 4.1 基于 logit 不确定性量化的股票预测算法所示。

算法 4.1: 基于 logit 不确定性量化的股票预测算法

输入：神经网络输出的 logit，类别标签 labels，超参数 q_1，q_2，u_1，u_2

输出：股票数据的不确定性分数

选择在训练过程中预测股票趋势正确的数据：

for each label in classes do

. If prediction_labels==train_labels

保存到 true_pred 数组中

设置超参数数值

使用高斯混合模型拟合到每个类别正确分类的样本的 logit 向量

选择最大的概率密度函数

计算股票数据的不确定性分数

4.3 趋势预测模型 SDENet

本章希望预测模型在进行预测时还能够具有考虑不确定性的能力，对于那些噪声数据、异常数据等不可靠的数据输入，模型应该尽量避免采用。

根据神经网络与动力学系统之间的联系，将神经网络中的正向传递可以看作是动态系统的状态转换，可以由神经网络参数化的常微分方程。但是，由于神经网络中的

微分方程是确定性的，无法捕获任何不确定性信息。所以，SDENet 考虑用随机微分方程表示了隐藏状态的转换，随机微分方程是在常微分方程的基础上加入噪声项所得，添加了布朗运动项以明确量化认知不确定性。随机微分方程的形式如下：

$$dX_t = \gamma(t, X_t)dt + \beta(t, X_t)dW_t \tag{4.7}$$

其中，X_t 为此方程的解，γ，β 表示两个函数，W_t 代表布朗运动。上式说明 X_t 满足对于任意的 $t(t \geq 0)$：

$$X_t = X_t + \int_0^t \gamma(\mu, X_\mu)d\mu + \int_0^t \beta(\mu, X_\mu)dW_\mu \tag{4.8}$$

SDENet 模型不仅能够对微分方程进行参数化以拟合预测函数的漂移网，而且包括对布朗运动进行参数化对训练分布之外的数据进行高扩散的扩散网。从控制的角度来看，漂移网控制模型达到良好的预测精度，而扩散网则表示随机环境下的模型不确定性。在深度学习的背景下，考虑到深度学习的输入可能具有数千个维度，而且本章节专注于监督学习和不确定性量化，因此选择使用具有固定步长的简单 Euler-Maruyama 进行有效的网络训练。在将时间间隔 $[0, T]$ 分为 N 个子间隔，通过以下方式模拟 SDENet:

$$x_{k+1} = x_k + f(x_k, t; \theta_f) \triangle t + g(x_0; \theta_g)\sqrt{\triangle t}Z_k \tag{4.9}$$

其中 Z_k 是标准高斯随机变量。求解 SDENet 的步数可以等效地视为传统神经网络定义中的层数，SDENet 的训练实际上是标准神经网络中的前向和后向传播，漂移神经网络 f 和扩散神经网络 g 交替优化。

神经网络作为确定性动力系统：SDENet 方法依赖于神经网络和动态系统之间的联系。当神经网络通过一系列隐藏层将输入 x 映射到输出 y 时，可以将隐藏层表示看作动力系统的状态。因此，可以通过用神经网络对其常微分方程进行参数化来定义动力学系统：

$$x_{t+1} = x_t + f(x_t, t) \tag{4.10}$$

x_t 代表第 t 层的隐藏状态。将这个方程重新排列为：

$$\lim_{t \to 0} \frac{x_{t+\triangle t} - x_t}{\triangle t} = f(x_t, t) \Leftrightarrow dxt = f(x_t, t)dt \tag{4.11}$$

神经网络常微分方程方法的思想是用神经网络参数化 $f(x_t, t)$，并利用常微分方程求解器在必要时评估隐藏单元状态。这种神经常微分方程可以以任意精度评估隐藏的单元动力学，并具有更好的内存和参数效率。

用布朗运动对认知不确定性进行建模：神经网络常微分方程是确定性模型，无法对认知不确定性进行建模。神经网络随机微分方程模型的核心是用布朗运动捕获认知不确定性。系统的连续时间动力学表示为:

$$dx_t = f(x_t, t)dt + g(x_t, t)dW_t \tag{4.12}$$

$g(x_t,t)$ 表示布朗运动的方差，表示动力系统的认知不确定性。该方差由系统处于哪个区域决定，如果系统处于训练数据丰富且认知不确定性低的区域，则布朗运动的方差将较小；如果系统处于训练数据稀少且认知不确定性高的区域，则布朗运动的方差将较大。

考虑不确定性估计的 SDENet：如上所述，可以使用布朗运动来了解认知不确定性。为了使系统能够达到良好的预测精度，同时得到可靠的不确定性估计，将 SDENet 模型设计为使用两个单独的神经网络来表示系统的漂移和扩散。SDENet 中的漂移网络 f 旨在控制系统以实现良好的预测精度。漂移网络 f 的另一个重要作用是捕获认知不确定性。这是通过将模型输出表示为概率分布来实现的。SDENet 中的扩散网 g 表示系统的扩散。

基于上述期望属性，本章提出以下目标函数来训练 SDENet 模型：

$$\min_{\theta_g} E_{x_0\ p_{train}} E(L(x_T)) + \min_{\theta_g} E_{x_0\ p_{train}} g(x_0; \theta_g) + \max_{\theta_g} E_{\tilde{x}_0\ p_{OOD}} g(\tilde{x}_0; \theta_g) \tag{4.13}$$

$$s.t. dxt = \overbrace{f(x_t, t; \theta_f)dt}^{driftnet} + \overbrace{g(x_0, t; \theta_g)dW_t}^{diffusionnet} \tag{4.14}$$

其中 $L(\cdot)$ 是依赖于任务的损失函数，例如用于分类的交叉熵损失，T 是随机过程的终端时间，P_{train} 是训练数据的分布，P_{OOD} 是分布外 (OOD) 数据。为了获得 OOD 数据，选择添加高斯噪声以获得噪声输入 x_0，然后根据卷积分布分配输入。与传统的每一层都有自己的参数的神经网络不同，SDENet 中的参数由每一层共享。这会减少参数的数量，并导致内存大幅减少。在目标函数中，还进行了简化，即扩散项的方差仅由起始点 x_0 而不是瞬时值 x_t 确定，这可以使优化过程更容易。

4.4 实验结果与分析

4.4.1 实验数据集和评价指标

股票交易者经常在推特等社交平台上发表一些个人对于股票市场发展变化趋势的见解。为了对模型的有效性进行验证，本文使用数据集[41]，该数据集是目前公认的用于股票价格预测的财务数据，反映了股票投资者对股票市场的看法。根据受欢迎的讨论程度，数据集选择了从 2014 年 1 月 1 日到 2016 年 1 月 1 日的时间段内，资本规模排名最高的 88 只股票，其中推文是根据股票交易者在推特上发布的股票的有效关键文本数据得出的。此外，2014 年 1 月 1 日到 2016 年 1 月 1 日期间的历史价格数据通过雅虎财经获得后纳入最终的数据集。该数据集包括两部分，分别是 88 支股票的价格数据和来自推特的评论数据。在本实验中仅考虑上涨和下跌两个股票趋势。上涨趋势的股票标签为 1，下跌的标签设置为 0。在该数据集中有一些具有较小波动率的实例，考虑到本章的目的是应用股票相关信息完成二分类任务，因此设置两个特定阈值-0.4% 和 0.5% 以去除移动百分比分布在两个阈值之间的实例。移动百分比为小于-0.4% 的实例标记为 0 (下降)，大于 5% 的实例标记为 1 (上升)。在划分数据集方面，将 2014 年 1 月 1 日到 2015 年 8 月 1 日之间 78% 的股票数据用于训练。2015 年 8 月 1 日到 2016 年 1 月 1 日期间的 22% 股票数据用于测试。按照以上原则进行划分，上涨和下跌两个类别的数据比例大致相等。最终训练集与测试集中的涨跌情况可以在表4.1中观察到。

表 4.1 类别统计

集合	时间	上涨类占比	下跌类占比
训练集	2014.01.01-2015.08.01	50.29%	49.71%
测试集	2015.08.02-2016.01.01	51.49%	48.45%

为了直观的表现模型的性能，采用的评价指标为：Accuracy、Precision、Recall、F1-Score[69]。

4.4.2 特征选择结果与分析

随机极端树在构建森林的过程中，对于股票的每个历史价格特征，使用信息增益率计算分割特征决策指标的归一化总缩减量，这个值称为特征重要性。依据特征重要性降序排列后，可根据需要选择前 k 个特征。

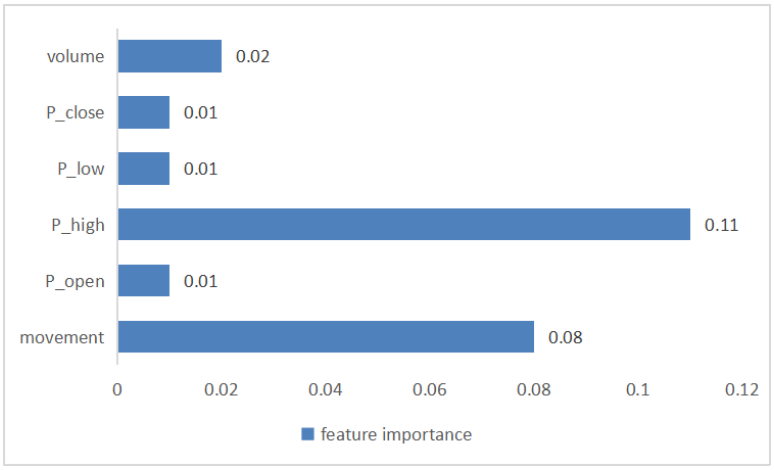

图 4.1 股票价格特征重要性对比图

根据极端随机树得到的特征重要性得分结果如图4.1所示，可以看出，历史价格特征的重要性降序后依次为最高价、涨跌幅、交易量、收盘价、最低价、开盘价。而且，最高价和涨跌幅二者的特征重要性分数明显高于其他的价格相关特征，交易量、收盘价以及最低价三个股票价格特征的重要性分数非常接近，开盘价的重要性分数最低。

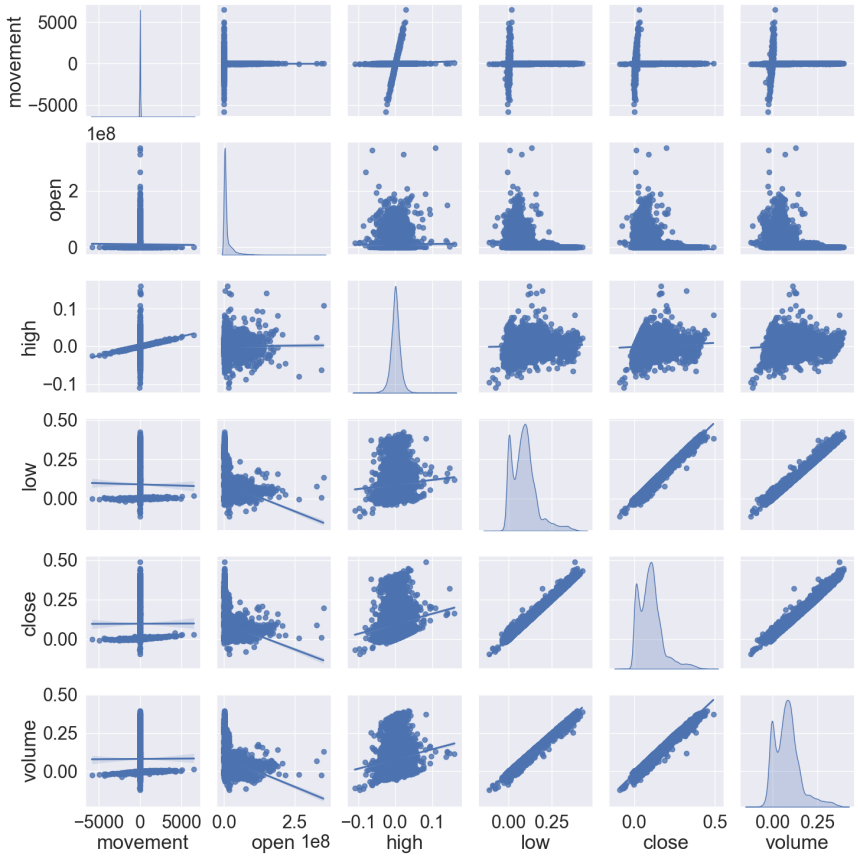

图 4.2 股票技术变量关系图

此外，本节使用 python 中的可视化库 Seaborn 展现了股票价格指标变量彼此之间的相关性关系，如图4.2所示。图中对角线位置为直方图分布。以第一行涨跌幅指标为例，可以看出涨跌幅与其余五个股票价格特征开盘价、收盘价、最高价、最低价、交易量存在线性相关关系。本节综合考虑了极端随机树的特征重要性得分与变量关系图中两两变量间关系，最终将涨跌幅作为后续预测的价格指标。

4.4.3 消融实验结果与分析

为了对 LogNet 中各部分的效果进行分析，本文将以下三个 LogNet 变体模型与 LogNet 模型进行消融实验，分析了各个部分对股票趋势预测的有效程度。

UAnalyst：此模型使用未经过任何处理的股票数据作为输入进行预测。

TreeAnalyst：为了证明特征选择部分的有效性，在此变体模型中去掉特征选择模块，但使用经过基于 logit 的模型处理的股票数据作为输入进行预测。

LogAnalyst：在此模型中，仅对股票的融合特征进行特征选择处理。

LogNet：加入特征选择模块以及经过 logit 模型处理后的股票预测模型。

表 4.2 变体模型性能情况

Methods	Extra Tree	Logit	Metrics			
			accuracy	precision	recall	F1-score
UAnalyst	-	-	50.521	0.71	0.51	0.35
TreeAnalyst	✓	-	50.963	0.51	0.51	0.45
LogAnalyst	-	✓	50.556	0.71	0.51	0.35
LogNet	✓	✓	64.556	0.70	0.65	0.65

表4.2描述了变体模型性能情况，根据表 4.2 进行可视化得到图4.3股票趋势三分类中的评价指标对比情况图可以看出：对股票数据不进行任何处理操作，将原始股票数据直接作为分类器的输入时，模型的准确率为 50.521%，接近于随机选择涨跌趋势。单纯的使用特征选择或者 logit 处理对股票进行预测的效果也并不理想。然而在模型中同时加入极度随机树用于特征选择以及加入基于 logit 模型对股票数据进行处理，相比于只使用二者中的单个部分，效果有了较大幅度的提升。这是因为股票数据具有低质量性，无效数据过多。可以看出，在输入分类器前，对股票数据进行处理，最终选择高质量、高信息量的数据进行预测将会提升股票预测模型的效果。

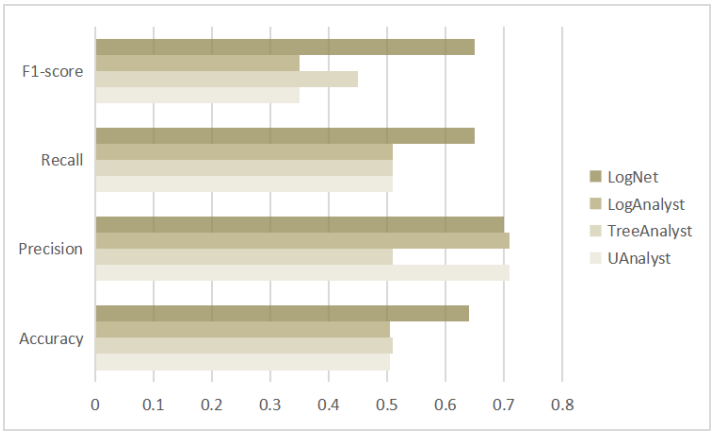

图 4.3 变体模型的指标对比情况图

4.4.4 对比实验结果与分析

为了证明本文提出模型的有效性，本节将所提出模型与以下现有的股票预测方法的性能进行对比：

多层感知机 (MLP)[65]：多个神经元层的神经网络分类器。

长短时记忆网络 (LSTM)[66]：能够处理特征序列问题的分类器。

双向长短时记忆网络 (Bi-LSTM)[67]：捕捉股票特征双向语义依赖的分类器。

多层注意力网络 (HAN)[39]：具有双层注意力机制的深度神经网络。

CredibleNet：第三章提出的基于多层注意力机制的股票趋势预测模型，采用特征融合以及三层注意力机制的深度神经网络。

LogNet：加入特征选择模块以及经过 logit 模型处理后的股票预测模型。

该部分讨论了上文提到的基线模型 MLP、LSTM、BiLSTM、HAN 与的方法模型 LogNet 在预测指标上的性能情况。

基线模型在二分类情况下的性能如表4.3所示，图4.4是依据表4.3所得到，展示了在股票预测二分类时各个基线模型的表现情况，根据实验结果可以看出在基线模型中，MLP 神经网络模型由于没有考虑到顺序上下文依赖的问题，预测效果一般，LSTM 与 BiLSTM 的性能结果对比说明了利用股票数据的双向语义信息进行预测，能够有效提升性能。CredibleNet 模型与传统的神经网络模型相比，具有更好的性能表现，而模型 LogNet 在准确性、召回率、F1 分数方面分别高于 CredibleNet 模型 0.966%、0.01%、0.02%、0.05%。根据上述 LogNet 与其他基线模型的性能比较可以看出，在大部分评价指标中 LogNet 优于其他基线方法。同时，可以看出 LogNet 在所有指标上均优于传

统的神经网络模型。对比结果表明本章的预测网络 LogNet 在预测股票市场未来趋势时的表现效果良好。

表 4.3 基线模型性能情况

Baseline Methodss	accuracy	precision	recall	F1-score	MCC
MLP	52.560	0.52	0.53	0.52	0.046
LSTM	52.926	0.53	0.53	0.53	0.059
Bi-LSTM	55.785	0.58	0.56	0.55	0.138
HAN(Hu et al.,2018)	58.344	0.71	0.58	0.53	0.283
CredibleNet	60.372	**0.74**	0.60	0.55	0.337
LogNet	**64.556**	0.70	**0.65**	**0.65**	**0.65**

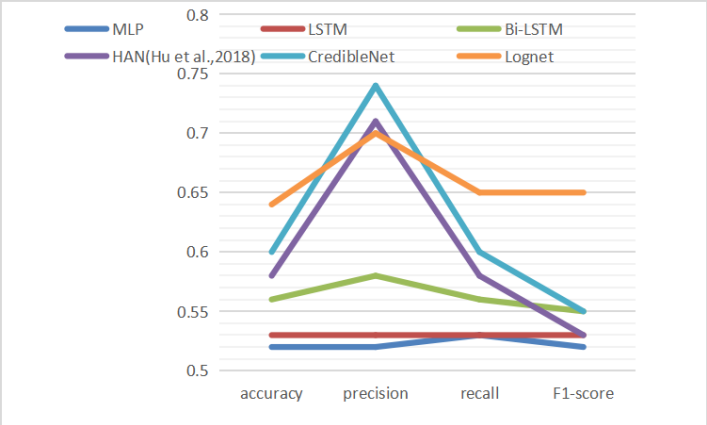

图 4.4 基线模型的指标对比情况图

4.5 本章小结

本章提出的 LogNet 模型首先使用极端决策树模型对股票的历史交易数据进行特征选择，选取对股票趋势预测具有较大影响的特征，提高模型预测效率。此外，本文将不确定性量化引入股票预测模型中处理股票数据信息，提出一种基于 logit 的不确定性量化的股票预测模型，利用同一类别的数据在 logit 输出上表现出相似性，利用高斯混合模型对数据源进行拟合，通过高斯概率密度函数对数据的不确定性进行量化并给出得分，基于此约束股票融合数据的特征，能够有效降低噪声数据的影响，提升模型泛化性能。最后对处理后的股票历史价格数据和文本数据进行特征融合，利用 SDENet 模型作为预测模型进行股票预测。

第 5 章　基于情感分析和 Transformer 的股票价格预测模型

5.1　任务分析与描述

本章工作的任务是将情感分析引入到股票预测任务中来，研究融合金融文本中的情感因素后对未来股票价格的影响。在推特文本数据上集成情绪分析模块，将市场中的公众情绪和历史股票价格据相关联后训练模型，增强股价预测模型性能。

从雅虎财经网站收集股票历史价格数据，原始数据由日期、开盘价、最高价、最低价、收盘价、在给定日期交易的股票数量以及涨跌幅七个变量组成。将原始数据中的七个变量作为历史价格的特征用于构建预测模型。在股票的技术指标中，收盘价是一天结束时的价格，又是第二天的开盘价，联系前后两天，因此最为重要。使用收盘价作为影响变量，根据股票收盘价的历史走势，预测未来的股票收盘价格。本章任务的目标是根据股票前期的历史价格数据以及文本情绪特征，预测股票的收盘价格。

关于股票相关的金融文本来自推特，使用自然语言工具包对推特上的股票文本情感进行计算和分析，使用已产生的价格和情绪得分来预测未来时间内股票的价格。在不同的 16 家公司进行测试，以评估模型并检查模型在各个领域的有效性。

5.2　研究方法描述

5.2.1　情感分析方法介绍

本文提出的基于情感分析的股价预测流程图如图5.1所示。将股票相关的文本数据使用 NLTK(Natural Language Toolkit, NLTK) 进行分词处理，NLTK 是自然语言处理工具包，常被用于搜索文本、词汇计数等工作，具有简易性、可扩展性且可模块化的优点。经过 NLTK 包处理后，使用 LM 金融情感词库中的 Positive(积极词频数表) 和 Negative(消极词频表)，以文本中积极词频数和消极词频数的占比作为文本数据的特征，式 (5.1)、(5.2) 计算 pos 和 neg 值作为非结构化文本数据的结构化特征。

$$pos = \frac{NumofPosWords}{TotalWords} \tag{5.1}$$

$$neg = \frac{NumofNegWords}{TotalWords} \tag{5.2}$$

以时间天为单位进行重采样，得出每日的 pos 和 neg 特征的平均值。并将得到股票的价格数据特征以及文本情感特征，按股票交易日期，对两者进行数据聚合以及归一化操作，扩大输入特征的范围。将融合投资者情绪以后的股票价格信息作为 Transformer 的输入，分析股票价格的下一步趋势。

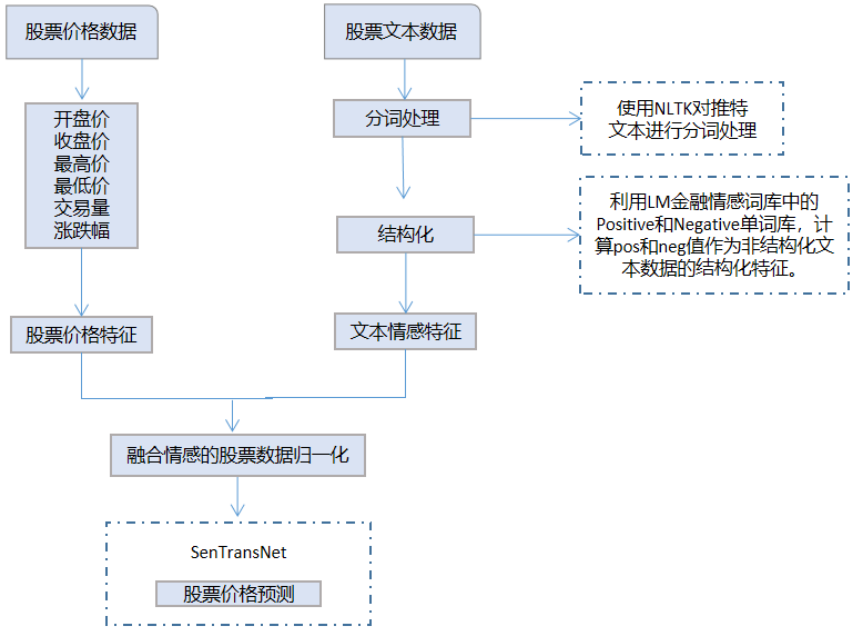

图 5.1 基于情感分析的股票预测模型流程图

5.2.2 SenTransNet 股票预测模型介绍

在 SenTransNet 股票预测模型，将通过情绪分析算法得到的股票文本与历史股票数据融合后的特征作为输入，以改进后的 Transformer 模型作为预测模型，对股票的未来收盘价进行预测。SenTransNet 模型结构如图5.2所示：传统的 Transformer 结构采用 Encode-Decode 体系结构。编码器由 M 个相同结构层堆叠组成，每一个结构层包括多头注意力层和完全连接层两个子层。在每个子层中使用残差连接和归一化来提高性能，编码器将输入序列的关键信息压缩成固定长度的向量。

位置编码：股票数据具有时间序列数据的特性，LSTM 本身是一种可以捕捉时序信息的顺序结构，包含了词语在句子中的位置信息。但使用自注意力机制时，词序信息就会丢失，模型无法了解单词在句子中的位置信息。位置编码通过生成的不同频率的正弦和余弦数据作为位置编码添加到输入序列中，从而使得模型可以捕捉输入变量的相对位置关系。

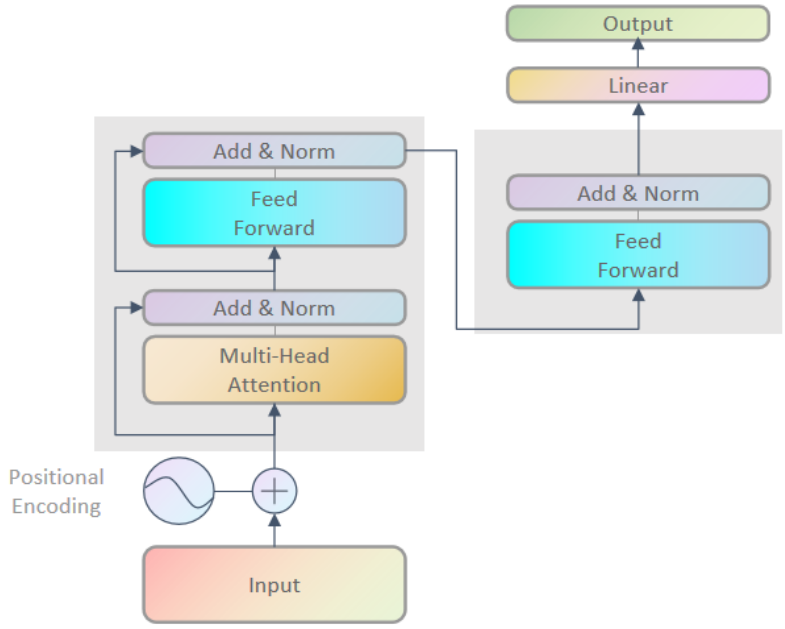

图 5.2 Transformer 模型结构图

多头注意力机制：自注意力机制首先创建 W_Q、W_K、W_V 矩阵将输入向量转换为计算多头自注意力值所需要的 Query(Q)、Key(K) 和 Value(V) 向量，然后使用 Q 向量与 K 向量计算点积，得到输入序列中的相关性得分 score。为了使训练时梯度能够更加稳定，使用 $\sqrt{d_k}$(d_k 表示 K 的维度) 对 score 进行处理。通过 softmax 函数对输出值进行归一化以获得概率分布，根据概率分布与 V 的乘积得到 $Attention(Q,K,V)$ 矩阵。输出矩阵公式如下所示：

$$Attention(Q,K,V) = softmax(\frac{QK_T}{\S\sqrt{d_k}})V \tag{5.3}$$

多层自注意力机制中，多个自注意力机制中的 Query(Q)、Key(K) 和 Value(V) 向量单独执行得到多个矩阵 $Attention(Q,K,V)$，将其拼接，通过线性层产生最终结果。多头注意力机制的公式如下：

$$MultiHead(Q,K,V) = Concat(head_1, head2, ..., head_h)W \tag{5.4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5.5}$$

其中，$i = 1, ..., h$，W_i^Q、W_i^K、W_i^V、W_i^O 为相应网络的权重，在本文中，考虑到本文

选取的每个股票数据集数量比极少，因此仅采用一层多头注意力机制，取 h=10。

增加残差块和归一化: 加入一个残差块，以避免深度神经网络在训练中发生退化问题。Norm 对数据数据进行归一化处理，有助于加快训练速度，提高训练稳定性。

前馈神经网络: 编码器中的每个子模型都包括全连接的前馈神经网络层 (FNN)，该层由 ReLU 激活函数和线性变换组成。先进行线性变换，将输入的多头注意力机制映射到更高维的空间中。然后经过 ReLU 非线性函数进行筛选。再进行线性变换变为原来的维度。前馈神经网络的公式如下：

$$FFN(x) = max(0, xw_1 + b_1)W_2 + b_2 \tag{5.6}$$

然后增加残差块并进行归一化处理，输入解码器中。解码器将编码器内容转换成输出，解码器的结构与编码器类似。

与原始解码器不同，不使用掩码注意机制，因为对于股价预测任务来说，解码器中的所有输入都是在没有未来信息的情况下观察到的历史数据。此外，考虑到股票预测任务的特殊性，在解码器中，并没有采用原模型中的解码器的结构，而是将解码器用了一个全连接层进行代替，用于输出股票预测值。

5.3 实验与结果分析

5.3.1 数据集描述以及超参数设置

为了确定 SenTransNet 模型进行股价预测的有效性，本文选取了服务行业、金融行业、工业品行业和技术四个行业的股票数据作为数据集[41] 对所提出的模型进行实验。时间为 2014 年 1 月 1 日到 2016 年 1 月 1 日，每一行业的股票数据集中有两个主要部分，其中包括推特文本数据集和历史价格数据集。在上述推特数据集中，其创建者已经使用 NLTK 包对所有推文进行了预处理，如标记化、主题标签、超链接和一些特殊的标识符也进行了处理。历史价格数据集是通过从雅虎金融公司提取历史价格来构建的，基于 4.1 章节的股票特征选择算法，选择涨跌幅、最高价以及收盘价作为股票的技术指标送到预测模型，以预测第二天的收盘价。

将数据集分为训练集和测试集两个部分。其中用于训练模型参数的训练集占总数据量的 80%。剩余的的 20% 数据作为测试集，用于评估模型的性能。每支股票的预测数据是每日收盘价。

在训练集上实施了大量实验，以预先确定最佳超参数。批处理训练的批处理大小设置为 64。使用均方误差作为损失函数，将预测值与实际值进行比较。学习率为 0.005 的 Adam 优化器用于训练模型。

5.3.2 评价指标

本章中选取的模型评价指标时均方根误差 (Root Mean Square Error, RMSE) 和平均绝对误差 (Mean Absolute Error, MAE)[70]。RMSE 用于表示预测值与真实值之间所存在的偏差，MAE 用于衡量预测值和真实值之间绝对误差的平均值。在验证过程中，RMSE 和 MAE 越小意味着模型的拟合效果越好。两者具体计算公式 (5.7)、(5.8) 如下所示：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5.7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5.8}$$

其中 y 为真实值，$\hat{y}$ 为预测值。

5.3.3 不同数据集下 SenTransNet 模型性能

不同领域和公司的股票在风险、波动性、盈利等方面都存在差异，使用不同领域多支股票数据集可以帮助验证模型是否具有更广泛的泛化能力，增加模型的可靠性。通过验证模型在不同领域不同公司股票上的表现，可以更好地评估模型的可靠性。为了验证本文模型的有效性以及鲁棒性，从金融业、工业、服务业以及科技行业四个不同领域中选取了四支个股作为数据集，验证 SenTransNet 模型在不同模型下的预测表现。在表5.1-表5.4中，使用 SenTransNet 模型对于不同的个股数据集进行情感分析，得到的积极情感句与消极情感句占总股票句数量的百分比，并分别展示在 positive 和 negative 数据列。SenTransNet 模型对各支股票的预测性能通过 RMSE 和 MAE 两个数据指标进行展现。

在金融行业中，考虑到行业地位与数据可获得性，选择了美国银行、摩根大通、花旗、万事达卡四家公司作为测试模型的数据集。这四家公司在金融行业中处于领先地位，它们作为金融数据集预测趋势具有代表性。其次，这些公司的数据通具有高度透明性和可靠性，它们会公布各种业务数据和指标，这些数据可以用于对公司的股票

进行分析和预测。

表 5.1 SenTransNet 在金融行业数据集下预测性能

Financial stock	positive	negative	RMSE	MAE
BankofAmerica(BAC)	26.23%	73.76%	0.051	0.033
Citigroup Inc(C)	26.97%	73.02%	0.179	0.083
JPMorganChase&Co.(JPM)	25.80%	74.19%	0.176	0.078
Mastercard Incorporated(MA)	54.48%	45.51%	0.194	0.082

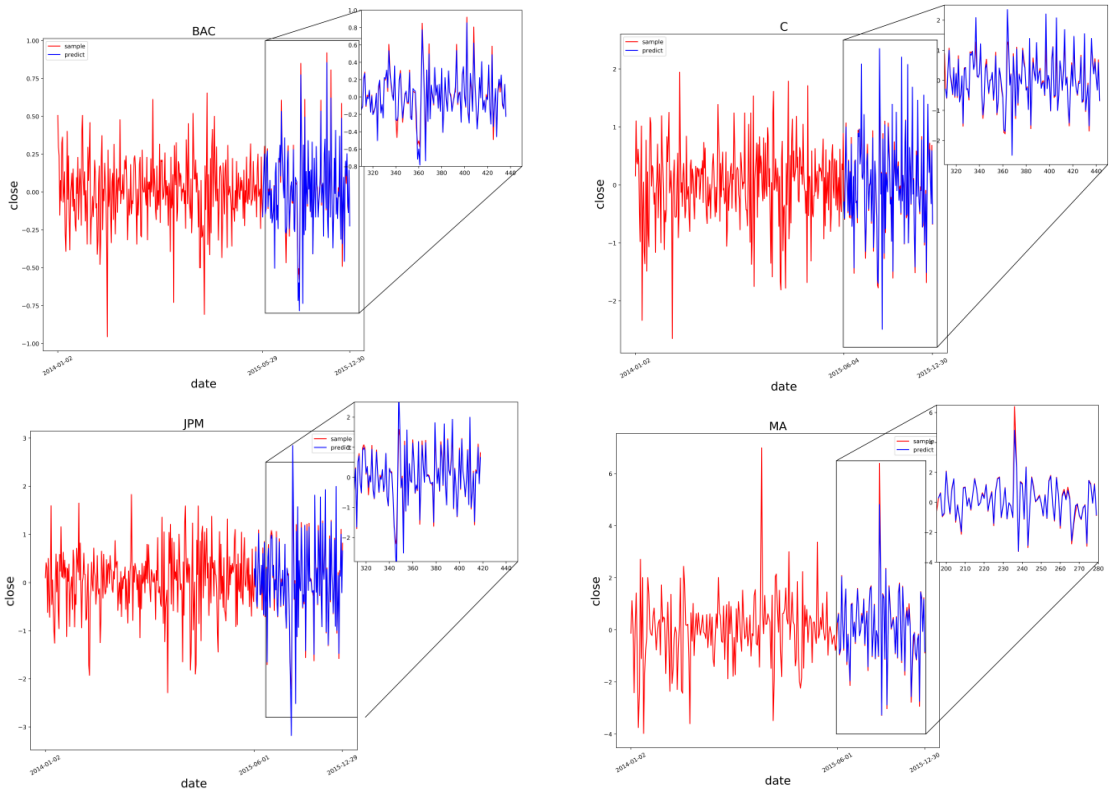

图 5.3 SenTransNet 对金融类公司预测拟合效果

表 5.2 SenTransNet 在工业数据集下预测性能

Industrial stock	positive	negative	RMSE	MAE
TheBoeing Company(BA)	42.44%	57.55%	0.139	0.092
Caterpillar Inc(CAT)	36.23%	63.76%	0.140	0.073
GeneralElectric Company（(E)	40.39%	59.60%	0.072	0.051
LockheedMartin Corporation(LMT)	52.32%	47.67%	0.356	0.199

在工业股票中，选取波音公司、卡特彼勒公司、通用电气公司、洛克希德马丁公司作为数据集。这些公司业务广泛、规模较大，在全球范围内都有着重要的市场地位

和影响力，经常成为媒体和分析师关注的焦点。此外，由于这些公司是公开上市公司，因此它们的财务报表和其他相关信息比较容易获得，有利于更好的分析股票趋势。

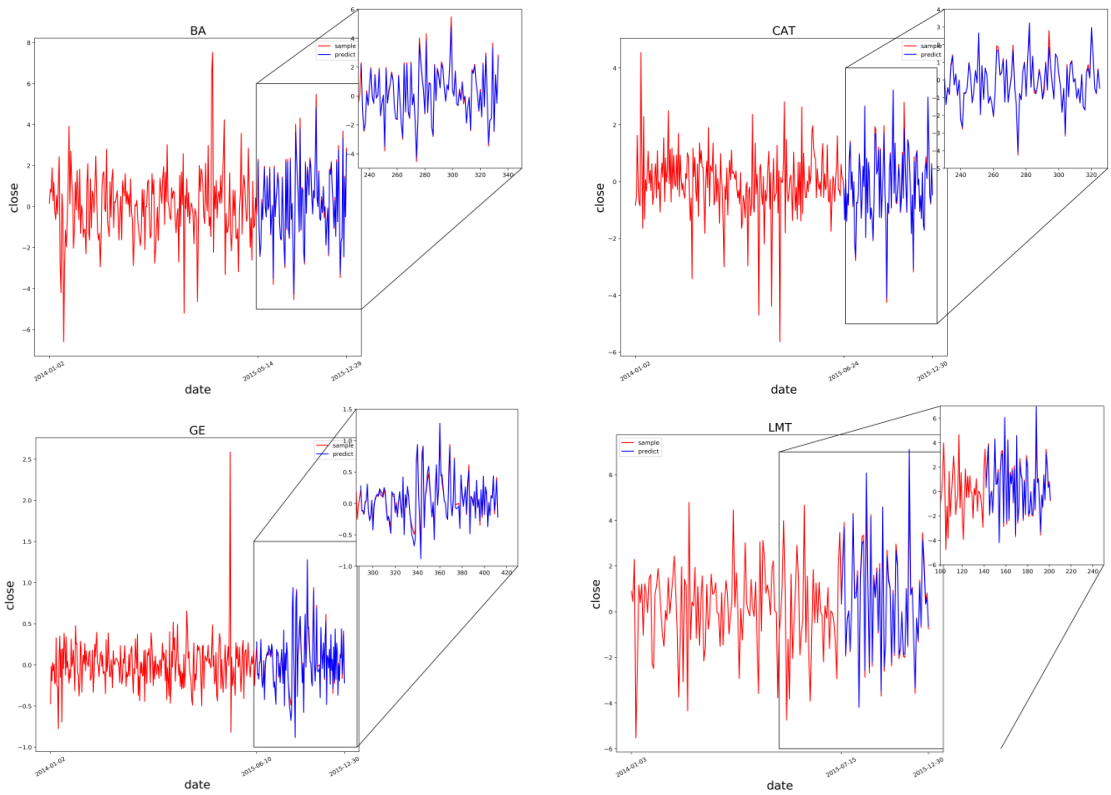

图 5.4 SenTransNet 对工业类公司预测拟合效果

亚马逊公司、康卡斯特公司、华特迪士尼、家得宝四家公司在服务业内有着重要的地位和影响力。市场地位稳固，吸引了众多股票投资者的广泛关注和讨论，因此，对它们股票表现进行预测具有重要意义。

表 5.3 SenTransNet 在服务业数据集下预测性能

Service stock	positive	negative	RMSE	MAE
Amazon.com Inc(AMZN)	48.02%	51.97%	2.290	0.771
Comcast Corporation(CMCSA)	50%	50%	0.051	0.038
TheWaltDisney（(DIS）)	35.35%	64.64%	0.808	0.245
The Home Depot(HD)	40.52%	59.47%	0.177	0.111

为了更加多元化的考虑到各种公司类型，本文选取思科系统公司、脸书公司、字母公司、英特尔公司四支科技股验证 SenTransNet 模型的预测效果。这四家公司代表了不同类型的科技公司，涵盖了网络设备、社交媒体、互联网搜索和芯片制造等不同领域。因此，通过分析这些公司的股票表现，可以了解 SenTransNet 预测模型对于科

技行业股票个股的预测性能。

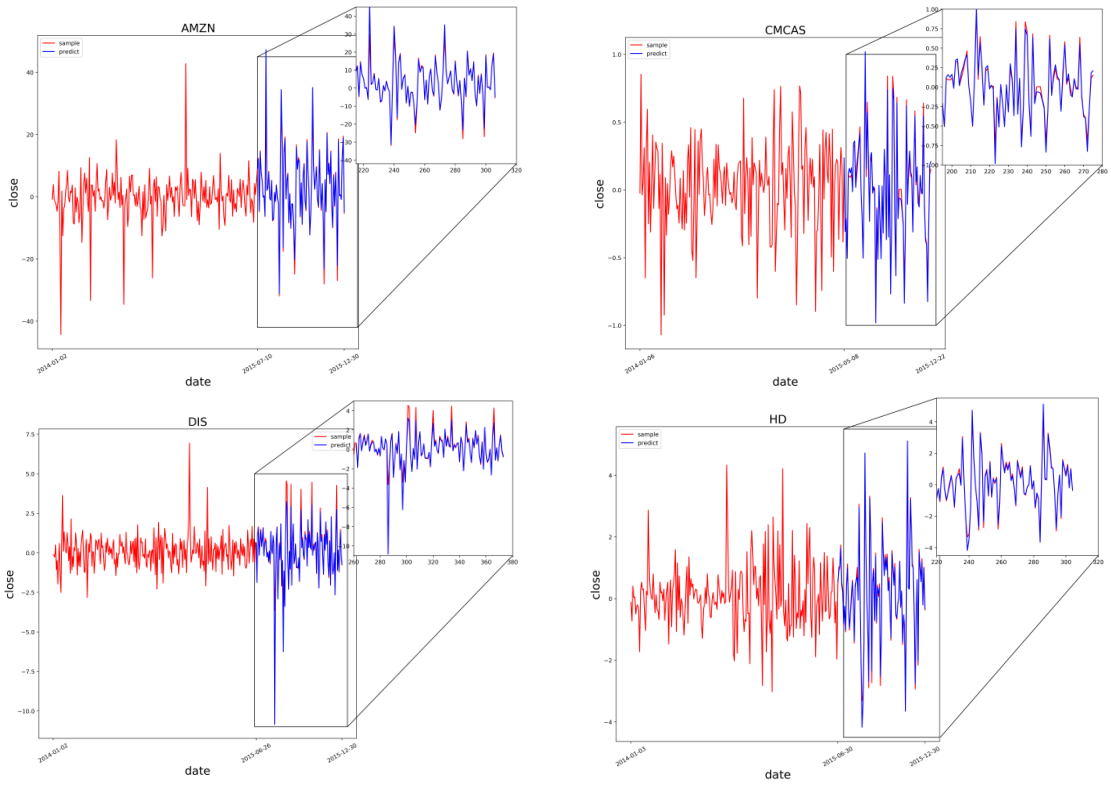

图 5.5 SenTransNet 对服务类公司预测拟合效果

表 5.4 SenTransNet 在科技类数据集下预测性能

Technology stock	positive	negative	RMSE	MAE
CiscoSystems Inc.(CSCO)	48.40%	51.59%	0.079	0.043
Facebook Inc.(FB)	16.18%	83.81%	0.109	0.073
AlphabetInc.(GOOG)	18.23%	81.76%	6.579	1.268
IntelCorporation(INTC)	33.76%	66.23%	0.071	0.049

此外，为了能够更加直观的对 SenTransNet 模型的预测效果有所了解，在图 5.3-图 5.6 中提供了 4 个行业 16 支股票的预测结果的拟合情况。通过局部放大图展示了 SenTransNet 模型对不同行业不同股票的预测效果，红线表示实际的股票收盘价，蓝色线表示使用 SenTransNet 模型预测的股票收盘价。从实验结果图可以看出 SenTransNet 模型对于不同行业的股票公司进行预测均可以实现较好的拟合效果，体现了本文所提出模型的稳定性和鲁棒性。

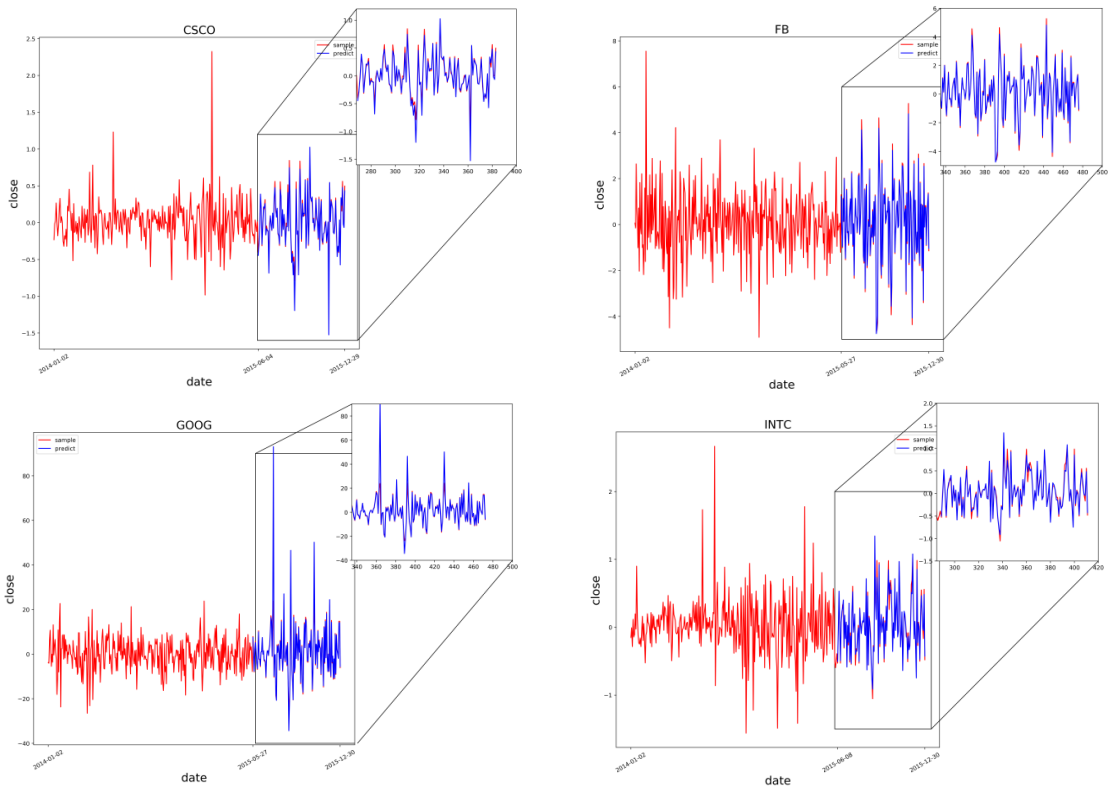

图 5.6 SenTransNet 对科技类公司预测拟合效果

5.3.4 对比实验结果与分析

为了验证 SenTransNet 模型的预测效果，本节以不同行业中的不同个股作为数据集，将 SenTransNet 模型与下述四个基线模型进行比较：

SenMLPNet[71]：输入采用股票的历史价格指标以及股票文本的情感分析增强模型性能。模型的决策层选择全连接网络。

SenLSTMNet[72]：该模型输入与 SenMLPNet 相同，采用 LSTM 模型对股票价格进行预测。

TransNet[73]：采用股票的历史价格指标作为模型的输入，不考虑股票文本的情感分析增强模型性能。决策模型层选择改进后的 Transfomer 模型。

SenTransNet：相比于 TransNet，输入中增加了股票文本的情绪因素，预测模型使用改进后 Transformer 模型。

表5.5显示了在 16 个数据集上四种模型的 RMSE 和 MAE 结果，展现了不同模型在不同的个股数据集下的预测表现。从 SenMLPNet 和 SenLSTMNet 方法的对比结果可以看出，在 16 个股票个股数据中的 10 支个股数据集上，SenLSTMNet 优于

表 5.5 不同行业数据集下各模型预测性能

	SenMLPNet		SenLSTMNet		TransNet		SenTransNet	
Metric	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
BAC	0.299	0.225	0.288	0.217	0.269	0.205	**0.051**	**0.033**
C	0.869	0.647	0.853	0.633	**0.096**	**0.050**	0.179	0.083
JPM	1.027	0.762	1.003	0.748	0.202	0.128	**0.176**	**0.078**
MA	1.418	1.080	1.432	1.096	**0.103**	**0.074**	0.194	0.082
BA	1.783	1.428	1.799	1.441	0.338	0.290	**0.139**	**0.092**
CAT	1.247	0.981	1.215	0.955	**0.113**	**0.071**	0.140	0.073
GE	0.390	0.298	0.369	0.280	0.143	0.099	**0.072**	**0.051**
LMT	2.222	1.804	2.231	1.814	0.510	0.302	**0.356**	**0.199**
AMZN	10.759	8.504	10.848	8.580	3.057	1.440	**2.290**	**0.771**
CMCAS	0.399	0.310	0.393	0.305	0.375	0.289	**0.051**	**0.038**
DIS	1.413	1.124	1.345	1.070	0.870	0.295	**0.808**	**0.245**
HD	1.530	1.168	1.506	1.148	0.411	0.175	**0.177**	**0.111**
CSCO	0.412	0.305	0.413	0.308	0.252	0.200	**0.079**	**0.043**
FB	1.786	1.399	1.753	1.369	0.205	0.175	**0.109**	**0.073**
GOOG	11.347	8.195	12.303	8.747	7.070	1.959	**6.579**	**1.268**
INTC	0.446	0.353	0.431	0.341	0.070	0.055	**0.071**	**0.049**

SenMLPNet，这是由于 MLP 对于股票价格这种具有强波动性的时间序列数据的拟合能力比较差，所以相比于具有时间序列建模能力的 LSTM 模型，预测性能表现要差。

采用 Transformer 作为股票预测模型的 TransNet 和 SenTransNet 方法的性能在 16 个数据集上均优于使用传统的神经网络 LSTM 和 MLP，表明了 Transformer 进行股票预测的有效性。通过 TransNet 和 SenTransNet 在两个评价指标上的结果展示，可以看出 SenTransNet 在大部分数据集中的股价预测性能要优于不使用情感分析模块的 TransNet 模型。综上所述，SenTransNet 的各项评价指标表现最优，这是由于 SenTransNet 考虑了更多的股票相关特征且具有更为强大的网络学习能力，因此其 MAE 和 RMSE 最小。

5.4 本章小结

为了进一步增强股票价格预测任务的预测精度，本章提出了一种基于基于情感分析和 Transformer 的股票预测模型 SenTransNet。该模型首先将情感分析引入到股票预测模型中，在训练和测试期间对文本数据进行情感分析。将分类后的情绪与股票相关的历史价格输入到股票模型中，进而预测股票接下来的价格情况。相比于传统的神经网络模型，Transformer 模型具有更优的特征提取能力、泛化能力以及运行效率，因此本章选取改进后的 Transformer 模型作为预测股价的决策模型，以股票的历史价格指标和非结构化数据情感分析特征为输入，对股票的收盘价指标进行预测。最后，选取不同行业的 16 个不同的股票公司进行测试，实验结果表明了 SenTransNet 模型进行股价预测的有效性。

第 6 章　总结与展望

6.1　本文工作总结

股票的趋势预测问题作为学术界和工业界的热点问题，吸引着众多相关工作研究者对此进行探索。传统的预测股票趋势的方法是对股票市场中的历史价格和交易量进行技术分析，然而这种方法对于描述股票市场剧烈变化具有较大的限制性。此外，另一种基本方法侧重于分析每个公司的财务报表，但是该方法无法捕捉近期趋势的影响。近些年，互联网的高速发展带动了社交网络媒体的发展，股票用户可以通过各种媒体平台自由发表对于股票市场的看法，来自网络媒体的文本内容成为股票投资者了解股票市场趋势和波动重要来源。此外，自然语言处理技术的飞跃式发展以及计算能力的大幅提升为处理和挖掘这些海量非结构化的文本信息提供了良好条件，通过自动分析股票相关文章等方式来进行股票趋势预测，有助于解决传统方法中仅以股票历史价格时序数据为影响变量的股票趋势预测问题。

在第三章中，提出了一种基于异构数据和多层注意力机制的股票价格预测模型，主要工作如下：面对股票预测问题，传统方法虽然在一定程度上实现了预测目标，但过于依赖指标的选择，缺乏全面描述股票波动的能力。对此，本文考虑融合异构数据信息，将相关的股票历史交易价格数据与非结构化的股票文本信息进行信息融合。对于股票相关的推特评论，本文使用 Glove 对其进行向量化，然后使用单词级注意力机制对一条评论中具有高价值的单词赋予高权重，使用句子级注意力机制对某天内所有的股票相关评论中的关键句子级评论赋予较高权重。对于股票的历史交易数据中的开盘价、收盘价以及涨跌幅三个特征进行信息融合得到时序特征，然后使用 BiLSTM 模型对该时序特征进行捕获时序特征。然后将非结构化文本数据与结构化历史价格数据进行特征融合，对于融合以后的特征利用 BiLSTM 模型进行编码，考虑到一天中存在多条评论消息，通过时间级注意力机制来获取天级新闻向量表示，以区分不同日期内股票评论的不同重要性，然后将处理后的数据加入到之后的 LSTM 模型中，输出下一步的股票趋势。通过对比试验和消融实验验证其有效性。实验证明，本文提出的模型 CredibleNet 能有效的提升股票预测性能。

在第四章中，本文提出了一种基于特征选择和不确定性量化的股票价格预测模型。主要工作如下：除了利用深度神经网络对股票预测模型的相关数据源进行处理

外，还可以从数据的不确定性方面入手。利用极端决策树方法从开盘价、最高价、交易量、涨跌幅等股票历史价格特征中选取关键特征。对于股票的评论文本信息，使用时间级注意力机制捕获天级评论文本表示，将两者进行特征融合。对于融合数据，本文使用了一种基于 logit 的不确定性量化方法：获取在训练过程中能够正确预测的数据元素，依据属于同一类别的数据在 logit 上应该具有相似性的理论，通过对正确预测的数据利用混合高斯模型进行建模，利用概率密度函数转化为不确定性分数。利用该不确定性量化方法，对预测数据给出不确定性得分，然后通过选取低不确定性分数的数据输入后续模型进行股票趋势预测，进而提高股票趋势的预测精度。在选取预测模型方面，本文采用了在预测的基础上还能考虑不确定性的模型 SDENet。该模型由两个神经网络 drift net 和 diffusion net 构成。通过对比实验与消融实验证明了该方法优于传统的股票预测模型。

在第五章中，提出一种基于情感分析和 Transformer 的股票预测模型，在该项研究中，将股票市场中投资者的情绪因素加入到股票预测任务中，本文使用自然语言处理库对社交媒体中股票相关的非结构化数据进行情感分析，获得情绪分数。然后将其与股票的技术指标进行融合，该融合特征作为预测模型的输入。在预测模型的选择上，本文并没有选取传统的机器学习模型，而是选择了高效、简便的 Transformer 模型。相比于其他模型，Transformer 模型能够更好的处理时序数据，挖掘其中包含的重要特征。本模型不仅使用了股票的历史价格数据，还结合当下网络社交媒体的兴起背景，还加入股民的情感特征进行信息增强。本文在多个不同行业的多支个股数据集上进行实验，结果表明，Transformer 模型具有较好的稳定性，且能够有效的提高股票价格预测任务的精度。

6.2 研究展望

本文提出了基于异构数据和多层注意力机制的股票预测模型、基于特征选择和不确定性量化的股票预测网络以及基于情感分析和 Transformer 的股票预测模型，提高了股票可用数据源的可用性，提升了股票预测结果的性能和可靠性。但本文的研究工作仍有很多不足之处，目前还存在以下方面可以进行进一步改进：

(1) 现有的股票预测任务对实时数据处理方面存在不足，现有的股票预测模型一般只能对历史数据进行预测，无法实时地处理实时股票数据，对于那些需要及时决策的股票交易者来说，这是一个亟待解决的问题。

(2) 影响股票趋势的因素多种多样，除了历史价格数据、社交媒体的股票评论和投资者情绪外，下一步的工作将考虑结合金融方面相关知识，研究影响股票趋势的其他相关因素，进行分析和特征融合，尽可能的考虑影响股票市场的相关因素，进一步的提高股票预测模型的可靠程度。

(3) 各大公司间的关系也是影响股票市场趋势的重要因素，如何将复杂多样的公司关系进行量化并加入到股票预测模型中，仍然需要进一步进行探究。

参考文献

[1] Picasso A, Merello S, Ma Y, et al. Technical analysis and sentiment embeddings for market trend prediction [J]. Expert Systems with Applications, 2019, 135: 60–70.

[2] Wafi A S, Hassan H, Mabrouk A. Fundamental analysis models in financial markets–review study [J]. Procedia economics and finance, 2015, 30: 939–947.

[3] Dong L, Wang Z, Xiong D. 基于文本信息的股票指数预测 [J]. Beijing Da Xue Xue Bao, 2017, 53 (2): 273–278.

[4] Bao W, Yue J, Rao Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory [J]. PloS one, 2017, 12 (7): 0180944.

[5] Hadavandi E, Shavandi H, Ghanbari A. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting [J]. Knowledge-Based Systems, 2010, 23 (8): 800–808.

[6] Zhang J, Teng Y-F, Chen W. Support vector regression with modified firefly algorithm for stock price forecasting [J]. Applied Intelligence, 2019, 49: 1658–1674.

[7] Yu P, Yan X. Stock price prediction based on deep neural networks [J]. Neural Computing and Applications, 2020, 32: 1609–1628.

[8] Chen Y, Hao Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction [J]. Expert Systems with Applications, 2017, 80: 340–355.

[9] Aguilar-Rivera R, Valenzuela-Rendón M, Rodríguez-Ortiz J. Genetic algorithms and Darwinian approaches in financial applications: A survey [J]. Expert Systems with Applications, 2015, 42 (21): 7684–7697.

[10] Nazário R T F, e Silva J L, Sobreiro V A, et al. A literature review of technical analysis on stock markets [J]. The Quarterly Review of Economics and Finance, 2017, 66: 115–126.

[11] Yang L, Xu Y, Ng T L J, et al. Leveraging BERT to improve the FEARS index for stock forecasting [C]. In The First Workshop on Financial Technology and Natural Language Processing, 2019.

[12] Li L, Leng S, Yang J, et al. Stock Market Autoregressive Dynamics: A Multinational Comparative Study with Quantile Regression [J]. Mathematical Problems in Engineering, 2016, 2016: 1–15.

[13] Nayak R K, Mishra D, Rath A K. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices [J]. Applied Soft Computing, 2015, 35: 670–680.

[14] Gao Q. Stock market forecasting using recurrent neural network [D]. University of Missouri–Columbia, 2016.

[15] Si J, Mukherjee A, Liu B, et al. Exploiting topic based twitter sentiment for stock prediction [C]. In

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013: 24–29.

[16] Idrees S M, Alam M A, Agarwal P. A prediction approach for stock market volatility based on time series data [J]. IEEE Access, 2019, 7: 17287–17298.

[17] Park C-H, Irwin S H. What do we know about the profitability of technical analysis? [J]. Journal of Economic surveys, 2007, 21 (4): 786–826.

[18] Nassirtoussi A K, Aghabozorgi S, Wah T Y, et al. Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment [J]. Expert Systems with Applications, 2015, 42 (1): 306–324.

[19] Wang W Y, Hua Z. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls [C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014: 1155–1165.

[20] Hagenau M, Liebmann M, Neumann D. Automated news reading: Stock price prediction based on financial news using context-capturing features [J]. Decision support systems, 2013, 55 (3): 685–697.

[21] Patel J, Shah S, Thakkar P, et al. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques [J]. Expert systems with applications, 2015, 42 (1): 259–268.

[22] Lu W, Li J, Li Y, et al. A CNN-LSTM-based model to forecast stock prices [J]. Complexity, 2020, 2020: 1–10.

[23] Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction [J]. Knowledge-Based Systems, 2019, 164: 163–173.

[24] Singh R, Srivastava S. Stock prediction using deep learning [J]. Multimedia Tools and Applications, 2017, 76: 18569–18584.

[25] Nosratabadi S, Mosavi A, Duan P, et al. Data science in economics: comprehensive review of advanced machine learning and deep learning methods [J]. Mathematics, 2020, 8 (10): 1799–1823.

[26] Akita R, Yoshihara A, Matsubara T, et al. Deep learning for stock prediction using numerical and textual information [C]. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016: 1–6.

[27] Yadav A, Jha C, Sharan A. Optimizing LSTM for time series prediction in Indian stock market [J]. Procedia Computer Science, 2020, 167: 2091–2100.

[28] Rather A M, Agarwal A, Sastry V. Recurrent neural network and a hybrid model for prediction of stock returns [J]. Expert Systems with Applications, 2015, 42 (6): 3234–3241.

[29] Rasheed J, Jamil A, Hameed A A, et al. Improving stock prediction accuracy using cnn and lstm [C]. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020: 1–5.

[30] Li L, Zhu F, Sun H, et al. Multi-source information fusion and deep-learning-based characteristics measurement for exploring the effects of peer engagement on stock price synchronicity [J]. Information Fusion, 2021, 69: 1–21.

[31] Li X, Xie H, Chen L, et al. News impact on stock price return via sentiment analysis [J]. Knowledge-Based Systems, 2014, 69: 14–23.

[32] Liu J, Lin H, Yang L, et al. Multi-element hierarchical attention capsule network for stock prediction [J]. IEEE Access, 2020, 8: 143114–143123.

[33] Li Q, Chen Y, Wang J, et al. Web media and stock markets: A survey and future directions from a big data perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30 (2): 381–399.

[34] Adam K, Marcet A, Nicolini J P. Stock market volatility and learning [J]. The Journal of finance, 2016, 71 (1): 33–82.

[35] Xing F Z, Cambria E, Welsch R E. Natural language based financial forecasting: a survey [J]. Artificial Intelligence Review, 2018, 50 (1): 49–73.

[36] Nassirtoussi A K, Aghabozorgi S, Wah T Y, et al. Text mining for market prediction: A systematic review [J]. Expert Systems with Applications, 2014, 41 (16): 7653–7670.

[37] Araci D. Finbert: Financial sentiment analysis with pre-trained language models [J]. arXiv preprint arXiv:1908.10063, 2019.

[38] Liu J, Lin H, Liu X, et al. Transformer-based capsule network for stock movement prediction [C]. In Proceedings of the first workshop on financial technology and natural language processing, 2019: 66–73.

[39] Hu Z, Liu W, Bian J, et al. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction [C]. In Proceedings of the eleventh ACM international conference on web search and data mining, 2018: 261–269.

[40] Huynh H D, Dang L M, Duong D. A new model for stock price movements prediction using deep neural network [C]. In Proceedings of the 8th International Symposium on Information and Communication Technology, 2017: 57–62.

[41] Xu Y, Cohen S B. Stock movement prediction from tweets and historical prices [C]. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018: 1970–1979.

[42] Tetlock P C. Giving content to investor sentiment: The role of media in the stock market [J]. The Journal of finance, 2007, 62 (3): 1139–1168.

[43] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market [J]. Journal of computational science, 2011, 2 (1): 1–8.

[44] Ranco G, Aleksovski D, Caldarelli G, et al. Investigating the relations between twitter sentiment and stock prices [J]. arXiv preprint arxiv:1506.02431, 2015.

[45] Rao T, Srivastava S, et al. Analyzing stock market movements using twitter sentiment analysis [J], 2012: 119–123.

[46] Heaton J, Polson N G, Witte J H. Deep learning in finance [J]. arXiv preprint arXiv:1602.06561, 2016.

[47] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J], 2017: 6000–6010.

[48] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 427–436.

[49] Chang J, Lan Z, Cheng C, et al. Data uncertainty learning in face recognition [C]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 5710–5719.

[50] 齐现英, 刘伯强, 徐建伟. 基于不确定性信息融合的高密度椒盐噪声降噪方法 [J]. 电子学报, 2016, 44 (4): 878–885.

[51] Dolezal J M, Srisuwananukorn A, Karpeyev D, et al. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology [J]. Nature communications, 2022, 13 (1): 6572–6585.

[52] Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges [J]. Information Fusion, 2021, 76: 243–297.

[53] Malinin A, Gales M. Uncertainty estimation in autoregressive structured prediction [J]. arXiv preprint arXiv:2002.07650, 2020.

[54] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning [C]. In international conference on machine learning, 2016: 1050–1059.

[55] Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural network [C]. In International conference on machine learning, 2015: 1613–1622.

[56] Wu H, Klabjan D. Logit-based uncertainty measure in classification [C]. In 2021 IEEE International Conference on Big Data (Big Data), 2021: 948–956.

[57] Kong L, Sun J, Zhang C. Sde-net: Equipping deep neural networks with uncertainty estimates [J].

arXiv preprint arXiv:2008.10546, 2020.

[58] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

[59] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014: 1532–1543.

[60] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees [J]. Machine learning, 2006, 63: 3–42.

[61] Xu H, Chai L, Luo Z, et al. Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices [J]. Neurocomputing, 2020, 418: 326–339.

[62] Feng F, Chen H, He X, et al. Enhancing stock movement prediction with adversarial training [J]. arXiv preprint arXiv:1810.09936, 2018.

[63] Zhang Q, Qin C, Zhang Y, et al. Transformer-based attention network for stock movement prediction [J]. Expert Systems with Applications, 2022, 202: 117239.

[64] Leng J, Liu W, Guo Q. Stock movement prediction model based on gated orthogonal recurrent units [J]. Intelligent Systems with Applications, 2022, 16: 200156.

[65] Gong S, Zhang D, Du S, et al. An Empirical Analysis and Research on the Prediction of Stock Trends Based on the MLP Neural Network Model [C]. In 2021 International Conference on Artificial Intelligence and Blockchain Technology (AIBT), 2021: 28–33.

[66] Yao S, Luo L, Peng H. High-frequency stock trend forecast using LSTM model [C]. In 2018 13th International Conference on Computer Science & Education (ICCSE), 2018: 1–4.

[67] 李见平. 基于双层 LSTM 模型的股票趋势预测研究 [J]. 科技与创新, 2021, No.175 (50-51).

[68] Dong L, Hong M, Huang M, et al. Logit-based stock prediction network [C/OL] // Zhong Y. In Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022), 2023: 125660M. https://doi.org/10.1117/12.2667702.

[69] Yao Y, Luo C, Leung K-C, et al. STGV-Similarity between trend generating vectors: A new sample weighting scheme for stock trend prediction using financial features of companies [J]. Expert Systems with Applications, 2023, 213: 119125.

[70] 王爱银 . 融合因果注意力 Transformer 模型的股价预测研究 [J]. 计算机工程与应用, 2023: 1–11.

[71] Turchenko V, Beraldi P, De Simone F, et al. Short-term stock price prediction using MLP in moving simulation mode [C]. In Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems, 2011: 666–671.

[72] Moghar A, Hamiche M. Stock market prediction using LSTM recurrent neural network [J]. Procedia Computer Science, 2020, 170: 1168–1173.

[73] Wang C, Chen Y, Zhang S, et al. Stock market index prediction using deep Transformer model [J]. Expert Systems with Applications, 2022, 208: 118128.