摘 要

证券市场是现代经济社会发展的重要组成部分，是解决资本供求矛盾、增加市场资金流动性和激活社会生产力的关键产物。

股票市场作为最活跃的证券市场，不断地受到媒体信息的影响。随着计算机技术的快速发展，媒体形式正由传统的新闻、社交媒体，逐渐转型为以信息交互为主的数字化互动媒体。然而，数字化互动媒体作为一种新兴的媒体形式，投资者与上市公司之间的相互关系缺乏深入的研究和探讨。除此之外，目前有关媒体对证券市场的影响性分析中，仍然主要关注社交媒体与新闻媒体，对于数字化互动媒体的研究非常有限。数字化互动媒体对证券市场是否存在影响？互动平台上的信息是否有助于提高证券市场波动分析的准确性？投资者与上市公司之间的相互关系该如何挖掘？这些问题都成为了极具科研价值的研究课题。

本研究基于深度学习视角，综合考虑数字化互动媒体的特性，在特征、决策、模型等方面做出了相应的贡献。在特征方面，本文不仅考虑了媒体视角，还聚焦于互动视角，探讨了数字化互动媒体中的内容特征、情感特征、显式交互特征和隐式交互特征，并提出了相应的预处理和提取方法。在决策方面，本文提出了一种基于三支决策的动态决策思想，以解决情感特征提取时效性问题。在模型方面，本文提出了一种基于注意力机制和双向结构的 Bi-JANETA 模型，可以有效捕捉数据的时间序列和特征信息。

最后，本研究基于数字化互动媒体数据，通过预测沪深 300 股指波动的准确性，检验本文方法的有效性。经过大量的对比实验证明，这些方法和技术的综合应用，能够提高股市预测的准确性和实用性。

关键词：数字化互动媒体；互动特征；深度神经网络；股票市场；

Abstract

Securities market is an important part of modern economic and social development, and a key product to solve the contradiction between capital supply and demand, increase market capital liquidity and activate social productivity.

The stock market, as the most active securities market, is constantly influenced by media information. With the rapid development of computer technology, media forms are gradually transforming from traditional news and social media to digital interactive media with information interaction as the main. However, as a new form of media, digital interactive media lacks in-depth research and discussion on the relationship between investors and listed companies. In addition, the current analysis of the impact of media on the securities market still mainly focuses on social media and news media, and the research on digital interactive media is very limited. Does digital interactive media have an impact on the securities market? Does the information on the interactive platform help to improve the accuracy of volatility analysis in securities markets? How to excavate the relationship between investors and listed companies? These problems have become research topics of great scientific value.

Based on the perspective of deep learning, this study comprehensively considers the characteristics of digital interactive media, and makes corresponding contributions in features, decision-making, models and other aspects. In terms of characteristics, this paper not only considers the media perspective, but also focuses on the interaction perspective, discusses the content features, emotion features, explicit interaction features and implicit interaction features in digital interactive media, and puts forward the corresponding preprocessing and extraction methods. In terms of decision-making, this paper proposes a dynamic decision-making idea based on three-way decision-making to solve the problem of timeliness of emotional feature extraction. In terms of model, this paper proposes a Bi-JANETA model based on attention mechanism and bidirectional structure, which can effectively capture

the time series and characteristic information of the data.

Finally, based on digital interactive media data, this study tests the effectiveness of the method in this paper by predicting the accuracy of the CSI 300 stock index volatility. Through a large number of comparative experiments, it is proved that the comprehensive application of these methods and techniques can improve the accuracy and practicality of stock market prediction.

Key words: digital interactive media; Interactive characteristics; Deep neural network; The stock market;

目 录

1.绪 论 1
1.1 研究背景及其意义 1
1.1.1 选题背景 1
1.1.2 研究意义 2
1.2 国内外研究现状 3
1.2.1 媒体对股市的影响 3
1.2.2 股市价格预测方法 4
1.2.3 文献述评 6
1.3 研究思路与框架 6
1.4 本文章节安排 8
2.相关理论知识 10
2.1 BERT 模型 10
2.2 深度学习模型 13
2.2.1 卷积神经网络 13
2.2.2 循环神经网络 14
2.2.3 长短期记忆网络 15
2.2.4 门控循环单元 17
2.2.5 JANET 18
2.3 注意力机制 19
2.4 三支决策基础理论 20
2.5 本章小结 22
3.基于数字化互动媒体驱动的股指波动预测 23
3.1 问题描述 23
3.2 问答文本数据获取及描述 26

3.3 数据特征处理 28
3.3.1 内容特征 28
3.3.2 显式交互特征 29
3.3.3 隐式交互特征 30
3.3.4 情感特征 33
3.4 Bi-JANETA 预测模型设计 37
3.5 本章小结 40
4.实验结果与分析 41
4.1 实验环境 41
4.2 数据集设置 42
4.3 实验设计 43
4.4 情感特征处理 45
4.4.1 分类质量比较 46
4.4.2 分类总成本比较 47
4.5 股指波动预测实验结果与分析 48
4.5.1 评价指标 48
4.5.2 基线模型 49
4.5.3 三支情感特征有效性对比 50
4.5.4 不同特征组合的预测性能比较 51
4.5.5 与基线模型预测性能比较 52
4.6 本章小结 53
6.总结与展望 54
6.1 论文总结和创新点 54
6.2 课题未来展望 55
参考文献 56
致　谢 61
在读期间科研成果目录 62

1.绪　论

1.1 研究背景及其意义

1.1.1 选题背景

证券市场的出现和发展是现代经济发展的重要产物，它为解决资本供求矛盾、增加市场资金流动性、激活社会生产力提供了有效的途径，从而推动了经济的快速发展。精准预测证券市场的波动趋势对于有效控制金融风险、指导投资者决策以及监管市场行为都具有极其重要的意义。

股票市场作为最活跃的证券市场，不断地受到信息的影响。股票市场上的交易会受到来自市场情绪效应、投资者非理性行为和各类事件所产生的信息干扰。Fama 在 1970 年提出了一种流行且最具争议的理论：有效市场假说(Efficient Markets Hypothesis，EMH)[1]，该理论认为在任何时间点，股票的市场价格都已经包含了该股票的全部信息。这意味着所有市场参与者在获取和利用信息方面具有同等的机会和能力，市场价格能够反映出股票的真实价值，不存在无风险套利机会，因此市场是高度有效的。根据这个理论，市场价格已经反映了所有相关的信息，所以任何形式的信息分析、技术分析、基本面分析等都无法获得超额收益。尽管该理论受到了不少批评，但它仍然是研究股票市场效率和信息有效性的基础理论之一，对投资者和市场参与者具有一定的指导意义。

现代行为金融学的研究表明，企业相关的外部信息，特别是媒体信息对证券市场波动的影响也具有非常重要的作用[4]。随着互联网技术的不断发展，媒体形式经历了从互联网初期门户网站为主导的“权威”新闻发布，到社交媒体为主导的“草根”自由信息发布的过程，再到如今的数字化互动媒体阶段。

在数字化互动媒体上，投资者可以直接向上市公司提出问题，了解企业的经营状况、业务发展战略和未来规划等重要信息。同时，上市公司也可以通

过数字化互动媒体回答投资者的问题，澄清不明确的信息和误解，以提高投资者对企业的了解和信任度。已经有大量的学者通过研究证实了互动媒体对资本市场的发展起到积极作用。比如，谭松涛[5]等研究发现，互联网信息平台的设立为投资者与上市公司管理层的沟通提供了便捷的渠道，增强了投资者获取信息的准确性，进而有效地提升了市场信息效率；丁慧等[6]通过对上证 e 互动的研究发现，互动媒体条件下投资者信息能力的提高能够显著降低股价崩盘风险。此外，互动平台的推出能够显著缩短投资者和上市公司高管之间的沟通距离，使得中小投资者更容易获得公司信息并积极参与公司治理。据中国互联网络信息中心（CNNIC）与中国证券管理委员会的统计数据显示，数字化互动媒体的注册用户已经达到亿级别，形成了强大的影响力，为投资者决策提供了重要参考，逐渐成为新一代影响证券市场波动的媒体因素[7]。

然而，目前现有的研究仍集中于新闻媒体、社交媒体阶段。与传统的新闻、社交媒体不同，数字化互动媒体中独特的“一问一答”机制，是投资者与上市公司之间的互动，会产生许多新的交互特征。例如表达内容、发布时间、问答情感的差异性，都可能会对最终的预测结果产生影响。因此，先前针对传统媒体的研究并不能直接运用于互动媒体阶段。继续探索和深入地了解数字化互动媒体，对于理解股票市场波动的形成机制以及提高股票市场预测的准确性具有重要的理论和实践意义。

1.1.2 研究意义

（1）理论意义

目前关于媒体信息对证券市场的研究主要集中于社交媒体、新闻媒体，对数字化互动媒体研究较少。本文不仅关注媒体视角，还聚焦于互动视角，在现有文献的基础上收集、整理并挖掘出有效的交互特征，并提出了新的股市波动预测模型。本研究的贡献丰富了相关文献的研究成果，并填补了相关研究领域的理论空白。此外，我们的研究还为未来相关研究提供了重要的学术参考和理论依据。

(2）现实意义

股票市场作为最活跃的证券市场，是经济运行的重要组成部分。股市的稳定运行对促进经济发展和优化资本结构配置具有重要意义。对股票市场波

动的预测不仅有助于对经济和金融的监测，而且还可以提供风险预警的功能，为稳健发展提供保障。本文面向股票市场，基于深度学习视角，针对数字化互动媒体的特性，在特征、决策、模型方面均做出了一定的贡献。这些方法的应用能够提高股票波动预测的准确性，比传统的金融预测方法更加科学准确。对投资者而言，可以帮助投资者客观地去分析市场行情，规避风险；对政府而言，可以帮助监管部门及时了解市场的走向和反应，从而更加准确地制定政策，提高政策的可行性和有效性。

1.2 国内外研究现状

1.2.1 媒体对股市的影响

经济学领域中的“理性经济人”假设通常被用来解释市场行为。然而，现实市场中存在的种种不确定性、信息不对称和其他因素使得这种假设难以全面解释市场中的行为现象。因此，许多学者开始将心理学的成果引入到股票市场的研究中进行分析得出投资者情绪会造成股价的波动。为了更好地理解这种现象，学者开始分析股民的评论、新闻报道和其他网络文本信息中隐藏的市场异动的根源。

在早期的统计学方法中，学者们将媒体信息数量作为特征指标，挖掘媒体信息与证券市场之间的关系[9,10]。Alanyali, Moat 和 Preis[11]利用《金融时报》中上市公司的每日提及次数发现，公司每日被提及次数与该公司股票的每日交易量均呈正相关。Preis, Moat 和 Stanley（2013）[12]通过分析与金融相关的 Google 搜索关键词的数量变化，为搜索关键词数量作为股市走势“预警信号”提供了实证支撑。在之后的时期，学者们开始尝试将信息量化为“词向量”以增强媒体信息对证券市场影响性的捕捉能力[13]。Wang, Huang 和 Wang（2012）[14]将媒体信息整体表征为特征向量，利用量化的媒体信息辅助证券市场波动解析，综合研究文本信息对证券市场指数的影响性。Akita et al.[15]将媒体信息转换为段落向量，用于证券市场的波动风险分析。

随着计算机技术的发展，学者们开始利用机器学习的方法，构建文本的情感指标，利用高级的自然语言处理方法给文章关键词汇或句子进行赋权，综合判别新闻文本的情感极性，并结合机器学习方法衡量证券市场的影响性

[16,17]。例如，Dickinson 和 Hu[18]将 N-gram 和“word2vec”文本表示技术与随机森林分类算法相结合，量化与股票市场相关推文的情绪值，证明了情绪值与公司股价之间的相关性。Yang, Mo 和 Liu[19]通过在 Twitter 领域内建立一个金融社区，使用一种市场情绪指数的加权算法来构建市场情绪指数，交叉验证的实验结果得出，由金融社区中的关键节点构建的加权情绪对金融市场的预测更为稳健。Li et al.[20]利用新闻中的名词集合来捕捉公司的基本面信息，并通过情感词汇来反映投资者的倾向性意见。另外，学者们通过情感语义方法来量化媒体信息的情感倾向。Bollen, Mao 和 Zeng[21]运用两种情绪跟踪工具（Opinion Finder、GPOMS）对句子的情感极性进行了识别以分析 Twitter 每日提要的文本内容，并分析是否与道琼斯工业平均指数（DJIA）的值相关。研究结果表明，加入特定的公众情绪维度可以显著提高 DJIA 涨跌预测的准确性，与其他研究相比，平均百分比误差减少了 6%以上。

值得注意的是，新型数字化互动媒体的兴起，带来了全新的问答交互模式，从而产生了更为复杂的互动特征，一些学者已经开始尝试解析投资者与上市公司的问答交互信息，对提问和回答的量化表征获得数字化互动媒体中交互行为指标。其中，丁慧,吕长江和陈运佳[6]在上证“e 互动”的研究中进一步发现，以互动指数度量的投资者信息能力的提升可以显著降低股价崩盘风险。张继勋和韩冬梅[22]检验了网络互动平台上公司管理层回复投资者提问的及时性和明确性对投资者投资决策的影响。研究发现公司管理层回复及时性和明确性是影响投资者投资决策的具体路径。郭培燕和李艳[23]以数字化互动媒体上的问答文本，从文本内容特征和问答结构特征出发，认为提问深度、情感倾向、回复时效性、回复信息量等是评价信息质量的重要依据。严炜炜、黄为和温馨[24]在对社交网络问答质量构建的答案质量评价体系中，包括了文本长度、关键词数量、句子数量、标点符号占比等。LiuB，FemgJ，Lium[25]研究发现文本中包含的通用词与停用词数量越少，质量越高。

1.2.2 股市价格预测方法

股票市场作为一种重要的证券市场，不仅反映了经济的状况，还对经济的发展产生了积极的影响。对于股票市场的预测在现代金融理论中具有重要

地位，因为准确地预测市场走势能够使投资者更好地了解市场趋势和投资机会，同时也能够为企业和政府的决策提供重要参考。

在早期，人们主要使用计量经济学模型预测股价的走势，常见的方法有自回归方法（AR）、移动平均模型（MA）、自回归移动平均模型（ARMA）和自回归综合移动平均（ARIMA）[26,27]。Pellegrini 等人[28]通过引入条件方差的广义自回归条件异方差（GARCH）模型，将 ARIMAGARCH 模型应用于金融时间序列的预测。尽管传统的计量模型在对符合其假设条件的时序数据进行拟合和预测时表现良好，但其函数形式通常较为固定和简化，并且它们往往依赖于一些严格的假设，例如独立同分布假设、T 分布等。然而真实的股市金融数据可能不完全满足这些假设，这导致传统计量模型的应用范围受到限制。

为了适应金融数据中非平稳性、非线性及高复杂性的数据特征，当下人们主要使用神经网络来解决复杂的预测工作，以得到更有效、更精确的结果。Chang（2012）[29]提出了一种连接神经网络（EPCNNs）以预测股票价格走势，并与 TSK 模糊系统、多元回归分析等其他模型的性能进行了比较，表明 EPCNN 可以对大多数数据提供非常准确的股票价格指数预测。Dai（2012）[30]将非线性独立成分分析（NLICA）和神经网络相结合来预测亚洲股票市场。Wang and Liu（2012）[31]改进了 Legendre 神经网络模型，并在预测模型中引入随机时间强度函数，为每个历史数据给出权重，以此来预测中国股市的价格波动。Kara 等人（2011）[32]使用了神经网络和支持向量机，预测每日伊斯坦布尔证券交易所(ISE)国家 100 指数，经过对比得出神经网络优于支持向量机。陈等人（2018）[33] 使用基于深度学习的股指期货预测模型、反向传播神经网络、极限学习机与径向基函数神经网络，对沪深 300 期货合约（IF1704）的交易数据进行测试，以评估它们对股票市场的性能。周等人（2018）[34] 将 LSTM 和 CNN 应用于股票市场的高频数据，采用滚动分割训练和测试集的方法来评估模型更新周期对模型性能的影响。

除了使用更先进的模型可以提升股票走势预测效果外，相关特征的选择，也是至关重要的一步。Vivek Sehgal 和 Charles Song 提出可以对网络上的文字载体进行情绪化研究，并以此得出股票走势预测结果[35]。Tsai 和 Hsiao [36]使用了主成分分析、遗传算法和决策树对多种特征进行组合，选出在预测股票方面有着重要的特征。Liu 和 Hu[37]对样本的不同特征赋予不同的权重值，以

提高股价预测的准确率。Nguyen 等[38]提取了社交媒体的情绪用于预测股票价格走势，与以往考虑整体的情绪不同，特定公司的情绪也被纳入股票预测模型。Picasso[39]等人考虑了技术指标和来自新闻文章的情绪的特征融合，并基于 20 只股票的投资组合预测股市趋势。Zhang[40]等人利用矩阵分解和张量分解相结合的方法，对新闻信息和社交媒体信息中提取的用户情绪进行融合。Kim[41]等人于 2019 年利用 LSTM 和基于卷积神经网络(CNN)的模型将时序特征和图形特征互相融合，并进行股价预测。

1.2.3 文献述评

上文总结了目前人们在金融科技领域做出的诸多探索与研究。首先从理论层面整理了媒体对股市的影响，接着介绍了深度学习在股票价格预测中的合理性与优越性，以及特征的选择对于预测性能提升的重要性。长久以来，大量的学者提出了众多有效的方法，不仅丰富了相关领域的理论，还极大地促进了股市预测研究的发展。然而，目前的相关研究仍然存在以下问题：

（1）针对互动媒体的研究较少。目前新闻、社交媒体对市场的影响已经有了大量的研究，然而，针对数字化互动媒体的研究才刚刚起步，缺乏相应的探索和深入的了解。

（2）互动媒体中交互特性还有待挖掘。目前已经有学者通过解析投资者与上市公司的问答交互信息来获取互动特征。然而，现有的研究还不够丰富，我们还有必要从数字化互动媒体信息中进一步挖掘出独特的互动特征。

（3）基于数字化互动媒体驱动的股指预测研究较少。股指通常可以反映股票市场的走势，对股指的波动进行准确预测，不仅可以帮助投资者规避风险，还可以帮助监管部门及时了解市场的走向和反应。然而，金融市场的预测是一个具有挑战性的问题。如何以数字化互动媒体数据作为驱动，设计合理的股指预测方法是一个极具价值的科研话题。

1.3 研究思路与框架

本文以数字化互动媒体上的问答文本数据为主要研究对象，使用深度学

习算法，以沪深 300 股指的价格涨跌预测为具体研究场景，提出了基于数字化互动媒体驱动的股价预测方法的研究框架，具备一定现实意义。具体而言，本研究主要内容有：

（1）在特征方面：本研究考虑到上市公司信息披露质量会影响投资者决策判断，于是从前人的研究中收集了用来评价问答质量的内容特征。此外，本文聚焦于数字化互动媒体中独特的互动视角，在进行广泛文献调研的基础上，综合考虑数字化互动媒体的特点，选取研究已证实会产生影响的显式交互特征，例如问答信息中问答长度、问答时间差等。此外，本文还提出了使用 CNN 提取问答信息矩阵中的隐式交互特征。

（2）在决策方面：本研究考虑到模型决策需要的准确性与有效性，于是在情感特征提取阶段引入三支决策，划分正负情感阈值，综合考虑决策代价，通过牺牲少量的准确度，使得模型的推理速度大大提升。

（3）在模型方面：本研究提出了一种基于 JANET 改进的 Bi-JANETA 模型，相比于 LSTM 模型更加轻量，记忆时序更短。Bi-JANETA 模型融入了双向机制，可以更好地结合特征序列的上下文信息。此外 Attention 机制的加入，可以让模型为不同特征分配权重，从而使得模型可以关注到更为重要的特征，提高预测效果。

本研究的论文结构如下所示：

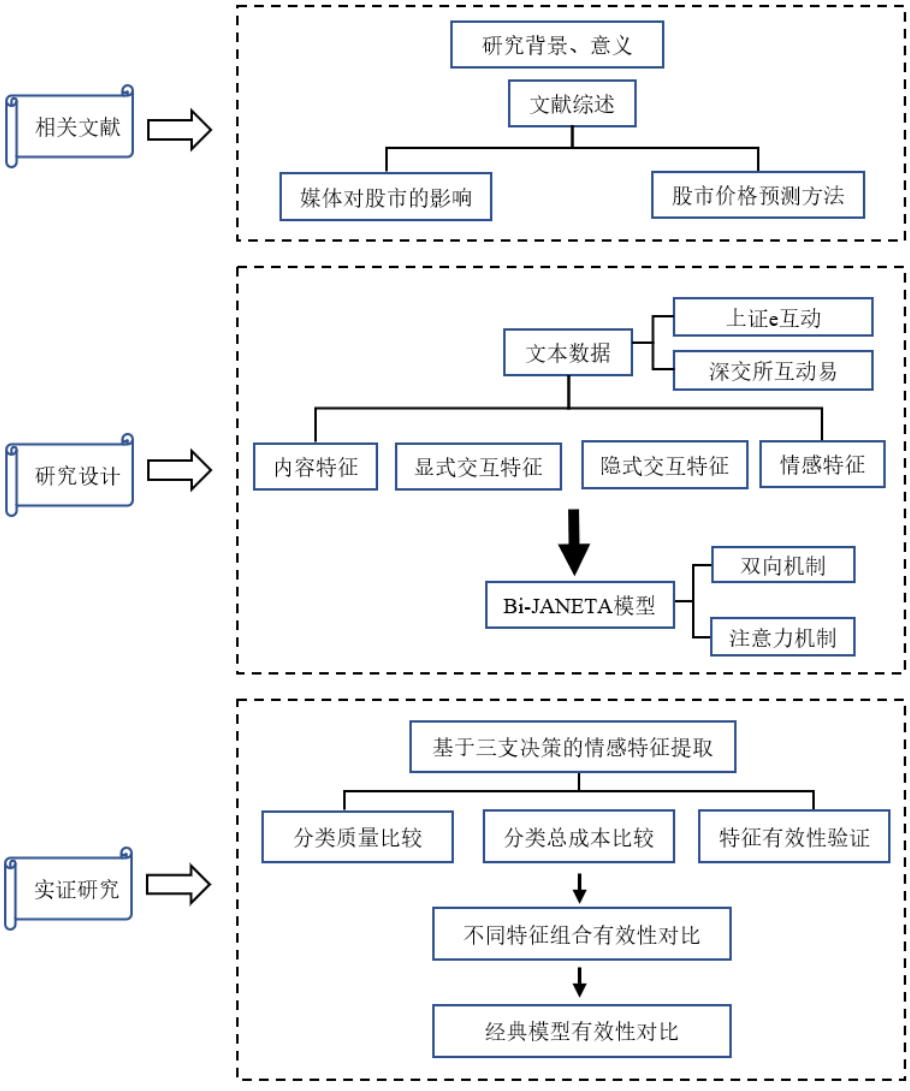

图 1-1 论文结构

1.4 本文章节安排

本文共有五章，现对每个章节进行简述，详情如下：

第一章的主要目的是介绍本文的研究背景和意义。本章介绍了证券市场在现代经济中的重要性，探讨了股票市场在国家经济金融活动中的作用，以及股价信息对投资者、企业和政府政策制定的重要性，说明了对股票市场进行预测的重要性。随后介绍了股价预测领域的相关研究现状，指出了现有的股价预测方法的局限性和挑战。最后，本章分析概括了文章的主要困难、挑战

和贡献，为后面章节的进行提供相应的支撑。

第二章主要分为三个部分，对本文所涉及的相关理论进行了研究。第一部分主要是介绍 BERT 模型；第二部分是深度学习模型，以及注意力机制的介绍；第三部分是三支决策的基础理论。本章可以为本文后续股指预测方法的设计和实现提供理论基础和方法指导。

第三章的主要内容为基于数字化互动媒体驱动的股指波动预测的设计与实现。本文首先聚焦于互动视角，提出了内容特征、显式交互特征等互动特征。其次使用三支决策及时、有效地提取情感特征。最后提出了 Bi-JANETA 预测模型。

第四章基于数字化互动媒体数据，以沪深 300 指数涨跌作为预测目标。首先通过实验确定了三支决策中的阈值，并根据实验证明了三支决策提取的情感特征能够兼顾有效性和及时性。然后通过不同特征之间的组合，证明了本文基于互动视角提出的互动特征的有效性。最后通过不同经典模型的对比，证明了本文提出的 Bi-JANETA 预测模型的有效性。

第五章是对本文工作的总结，以及课题未来的展望。

2.相关理论知识

2.1 BERT 模型

谷歌公司于 2018 年发布了一种基于深度学习的新型预训练模型 BERT[42]。BERT 使用了大量的无标注数据进行预训练，研究人员在此基础上针对特定自然语言处理任务作微调，便可应用于下游目标任务，主要模型结构如图 2-1 所示。

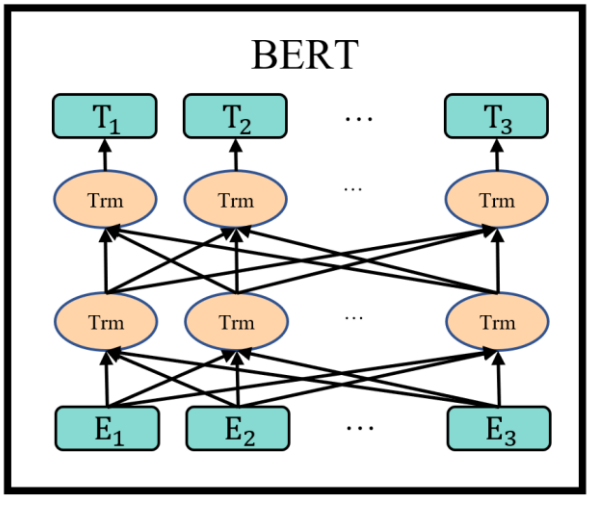

图 2-1 BERT 主要模型结构

图 2-1 中的 Trm 模块为 BERT 模型的核心部分，也就是 Transformer[43]中的 Encoder 模块，如图 2-2 所示。

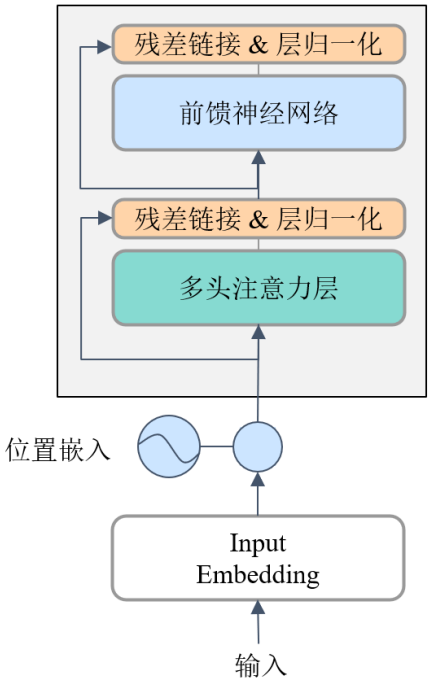

图 2-2 Encoder 结构

BERT 的输入如图 2-3 所示，由图可以直观地显示出输入的数据是由三种携带不同含义信息的向量进行相加，分别为词嵌入层(Token embedding)，段嵌入(Segment embedding)层和位置嵌入(Position embedding)层。

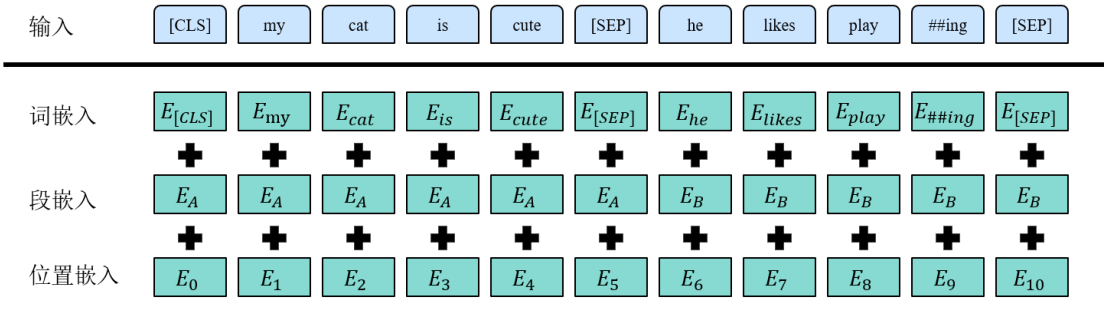

图 2-3 BERT 输入

在文本中，某个单词的意思可能会随着它在句子中的位置、上下文等信息改变而改变，这样的现象称为“一词多义”。在自然语言处理中，对于“一词多义”的问题，需要利用上下文信息来确定该词在特定上下文中的语义。为此，可以使用注意力机制来引入上下文信息，以便更好地理解文本。具体而言，自注意力机制来获得每个词在不同语义空间下的表示，然后使用多头注意力机制来将这些表示组合起来，以便更好地理解文本。多头注意力机制是

一种利用多个注意力机制来获得多个语义表示的方法。在多头注意力机制中，模型可以将不同的自注意力模块用于不同的语义空间，从而获得多个语义表示。然后，可以使用线性组合的方法来将这些语义表示组合起来，以获得一个更全面的理解。

采用多头注意力机制计算富含语义信息向量的具体步骤如下：令目标字向量为Q，目标字上下文各个字向量为K，其字向量维度为d_k，目标字及上下文字各自的原始向量为V，文本句长为m。首先计算Q、K之间的相似度，并将得到的值应用$softmax$函数得到权重值。并对得到的权重值进行加权求和，得到自注意力向量。随后计算每个Q、K、V中h个不同部分的$Attention$值，并将h个$Attention$结果拼接获得最终的向量。计算公式如下所示：

$$f(Q,K_i)=Q^TK_i,\qquad i=1,2,3\ldots \tag{2-1}$$

$$\alpha_i=softmax\left(\frac{f(Q,K_i)}{\sqrt{d_k}}\right),\qquad i=1,2,3\ldots \tag{2-2}$$

$$Attention(Q,K,V)=\sum_{i=1}^{m}\alpha_iV_i,\qquad i=1,2,3\ldots \tag{2-3}$$

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{2-4}$$

$$MultiHead(Q,K,V)=Concact(head_1,\ldots,head_h)W^0 \tag{2-5}$$

多头注意力方法可以综合各个位置上语义的信息。对于一个注意力头的情况，很多重要信息就会被平均计算给抑制。由于多头注意力机制所有头的维数总和与只有一个注意力头的维数相同，所以总的计算成本与全维数的单头注意力机制差不多。将多头注意力模块进行残差连接、标准化、线性转换等步骤，即可构成 Transformer 中的 Encoder 模块，也就是上文提到的 Trm 模块。经过多个 Transformer Encoder 模块的运算即可得到 Bert 模型的输出。

在使用 Bert 模型时，文本前后会插入两个特殊的符号，分别是[CLS]和[SEP]。其中，[CLS]符号并未携带明确的信息，它所对应的输出向量已经获得了整句文本的语义信息，在表达上会更加公平、准确，可以将其用于下游的分类任务。 [SEP]符号用于分开两个输入句子，适合用于其他任务，例如对话问答、序列标注等。总之，BERT 模型的[CLS]符号和[SEP]符号的插入和使用，旨在方便地使用 Bert 模型处理多种自然语言处理任务。

2.2 深度学习模型

2.2.1 卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）的应用主要包括图像处理、自然语言处理和音频处理等方面。在这些应用中，矩阵作为输入数据的基本形式，CNN 可以有效地提取输入数据的特征，从而实现对数据的分类、识别、分割等任务，已成为深度学习领域的重要工具。

CNN 的主要特点是卷积层和池化层的结构。卷积层的作用是对输入的矩阵进行卷积操作，提取图像的局部特征。卷积操作可以看作是对输入矩阵的滤波处理，通过不同大小的卷积核对矩阵进行扫描，提取出局部的特征信息。卷积操作的特点是局部连接和权值共享，即卷积核的参数在不同位置的卷积操作中共享，从而减少了参数数量，加速了模型训练。

池化层是卷积神经网络中另一种常见的层，用于减少特征图的尺寸和数量。池化层通常通过滑动窗口的方式对输入的特征图进行扫描，并将每个窗口内的特征值进行聚合，得到一个更小的输出值。最大池化是池化的一种常见方式，它选择每个窗口内的最大特征值作为输出值。平均池化则选择每个窗口内的特征值的平均值作为输出值。池化层的参数通常是窗口大小和步幅大小。窗口大小指的是池化层所使用的滑动窗口的大小，通常为正方形或矩形。步幅大小指的是滑动窗口每次移动的步长大小。通过调整这些参数，可以控制池化层输出特征图的大小和数量。

CNN 的最后一层通常是全连接层，将特征图转换为类别概率，进行分类或回归任务。在全连接层中，每个神经元都与上一层的所有神经元相连，权值矩阵的大小与上一层的神经元数量成正比，这使得全连接层的参数量非常大。为了避免过拟合和提高泛化性能，通常使用 dropout、正则化等方法对全连接层进行优化。

图像处理是 CNN 在矩阵处理领域应用最为广泛的领域之一。在图像处理任务中，图像可以看作是由像素点组成的二维矩阵。CNN 可以通过卷积层提取图像的特征，从而实现对图像的分类、目标检测、语义分割等任务。在图像分类任务中，CNN 通过多层卷积和池化操作，可以提取出图像的局部特征，

然后通过全连接层将这些特征映射到各个类别的概率上。例如，AlexNet[44]和VGG[45]等经典 CNN 模型在 ImageNet 数据集上取得了很好的分类效果。

在自然语言处理领域，CNN 主要应用于文本分类和情感分析任务。在这些任务中，输入数据可以表示为一个文本矩阵，每一行表示一个词向量，每一列表示一个特征。在文本分类任务中，CNN 可以通过多个卷积层提取出文本中的局部特征，然后通过池化层进行特征降维，最后通过全连接层将特征映射到不同的类别上。例如，Kalchbrenner 等人[46]提出的 Dynamic Convolutional Neural Network（DCNN）在情感分析任务中取得了很好的效果。

2.2.2 循环神经网络

循环神经网络（RNN）的设计是为了处理具有序列性质的数据，例如文本、音频和时间序列数据等。在这些数据中，每个输入与之前的输入是相关联的，并且存在一定的上下文关系。

循环神经网络的主要结构特点是它具有循环连接，使得网络能够对前面的输入进行记忆。这种记忆功能使得循环神经网络能够有效地处理序列数据，并对序列中的上下文信息进行建模。具体来说，循环神经网络中的隐藏状态可以通过对之前的输入和隐藏状态进行计算而得到，使得网络能够有效地捕捉到序列数据中的长期依赖关系。

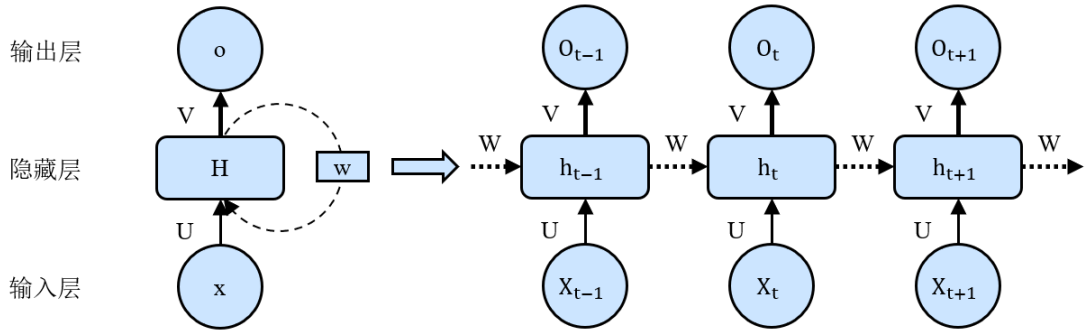

图 2-4 RNN 结构图

RNN 的模型结构如图 2-4 所示。图中左侧为 RNN 的总体结构，右侧为 RNN 的展开结构。通过右侧的展开图，我们可以清晰的看到，W 是隐藏层的权重矩阵，U 是输入层到隐藏层的权重矩阵，V 是隐藏层到输出层的权重矩阵，W 、U 、V 三者在任意时刻可以共享，W 在全程每一个时刻都是一样的。

t 时刻的隐藏单元 h_t接受前一时刻隐藏单元 h_{t-1} 的数据以及 t时刻的输入数据x_t， 输出为 o_t，而h_{t+1} 的输入为 h_t和x_{t+1} ，以此类推可以看出 RNN 结构中每一时刻的状态都依赖前一时刻的状态。RNN 网络计算公式如下所示：

$$ot = g(V \cdot h_t) \tag{2-6}$$

$$ot = g(V \cdot h_t)h_t = f(U \cdot x_t + W \cdot h_{t-1}) \tag{2-7}$$

其中的 f 、 g 为可选择的激活函数， x_t为 t 时刻的输入， o_t为 t 时刻的输出、h_t为 t 时刻隐藏单元的值，h_{t-1} 为前一时刻隐藏单元的值。

虽然循环神经网络（RNN）在处理序列数据方面具有一定的优势，但在实际应用中存在着一些问题。例如当序列非常长时，网络可能会出现梯度消失或梯度爆炸的问题，导致网络无法有效地学习长期依赖关系，这一缺陷也导致了它的记忆能力有限，很难有效地保存长期的信息。因此有学者在 RNN 的基础上提出了长短期记忆神经网络。

2.2.3 长短期记忆网络

长短期记忆网络（Long Short-Term Memory，LSTM）是一种循环神经网络的变体，它具有独特的结构设计，能够有效地解决 RNN 网络中长距离依赖的问题。相较于 RNN 网络，LSTM 通过增加一些门控单元来增强网络的记忆功能，从而使网络能够有效地处理长期依赖关系，并避免梯度消失或梯度爆炸的问题。图 2-5 显示了 LSTM 网络在时间步 t 的记忆模块的内部结构。

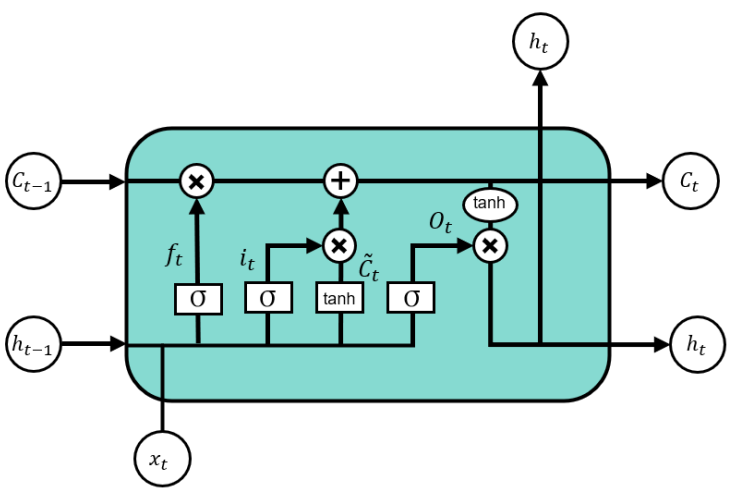

图 2-5 LSTM 结构图

在图 2-5 中，σ是 sigmoid 函数，h_t为隐藏状态，可以将其视为短期记忆；C_t为单元状态，可以将其视为长期记忆，C_{t-1}为旧的单元状态，x_t为当前输入的信息。

LSTM 能够灵活地控制记忆和遗忘，从而在长序列处理中表现出更好的效果。具体而言，LSTM 网络增加了一个称为“单元状态”的内部状态，并添加了三个门控制单元状态的流动，分别为遗忘门、输入门和输出门。

输入门和遗忘门都是控制信息流入或者流出记忆单元的门控机制。在每个时间步中，输入门和遗忘门会根据输入和上一个时间步的隐藏状态来计算一个 0 到 1 之间的值。输入门的值决定了新的输入是否应该被添加到记忆单元中，而遗忘门的值则决定了上一个时间步的记忆单元中哪些信息需要保留下来。具体而言，遗忘门的值越接近于 0，表示之前的信息需要被遗忘，而值越接近于 1，则表示需要保留。

在经过遗忘门和输入门之后，需要计算单元状态，并将更新的信息加入到单元状态中。这个计算过程基于上一个时间步的单元状态和输入门、遗忘门的值来更新当前时间步的单元状态。具体而言，LSTM 使用一个更新门通过运用 tanh 激活函数来将当前时间步的输入和上一个时间步的隐藏状态进行组合计算。这个更新门的值被乘以输入门的值，表示需要将新的信息添加到记忆单元中，同时也被乘以遗忘门的值，表示需要从上一个时间步的单元状态中保留一部分信息。这样就可以使用遗忘和保留的值来更新单元状态，从而保留对序列中重要信息的记忆。

最后，输出门决定了下一个隐藏状态的值，这个值是基于当前时间步的单元状态和输入门、遗忘门计算得到的。输出门控制着记忆单元中信息的流出，它通过计算一个 0 到 1 之间的值，确定了当前时间步的隐藏状态中哪些信息需要输出到下一层网络或用于其他任务。LSTM 网络的计算公式如（2-8）~（2-13）所示。

遗忘门f_t计算公式：

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2-8}$$

输入门 i_t 计算公式：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2-9}$$

候选值 $\tilde{c_t}$计算公式：

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (2-10)$$

更新单元状态 c_t计算公式：

$$\mathrm{C_t} = \mathrm{f_t} * \mathrm{C_{t-1}} + \mathrm{i_t} * \tilde{\mathrm{C}}_\mathrm{t} \quad (2-11)$$

输出门 o_t计算公式：

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2-12)$$

最终输出信息 h_t计算公式：

$$\mathrm{h_t} = \mathrm{o_t} * \tanh(\mathrm{C_t}) \quad (2-13)$$

其中W_f为遗忘门的权重矩阵， b_i表示输入门的偏置项，b_f 表示遗忘门的偏置项；W_i为输入门的权重矩阵； b_o为输出门的偏置项，W_c 为候选值的权重矩阵，b_c 为它的偏置项；W_o 为输出门的权重矩阵。

2.2.4 门控循环单元

门控循环单元（Gated Recurrent Unit，GRU）已成为一种广泛应用的序列模型，相比于长短时记忆网络（LSTM），GRU 网络的结构更为简单，但能够有效地解决 RNN 网络中存在的长距离依赖问题，尤其适用于数据量不大的语料库。GRU 将 LSTM 中的遗忘门和输出门结合为一个称为“更新门”的门控制单元状态的流动，相比 LSTM 的三个门。

在许多情况下，GRU 的性能几乎与 LSTM 相当。不仅如此，由于 GRU 具有更简单的结构和较少的参数量，因此通常在训练时需要的计算资源和时间也更少。图 2-6 显示了 GRU 的结构图。

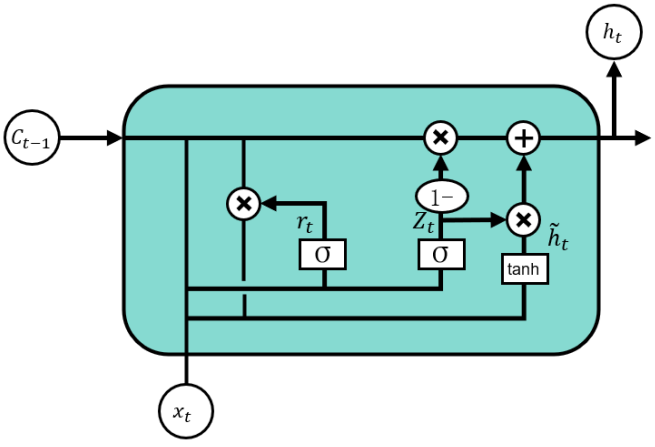

图 2-6 GRU 结构图

GRU 具有两个门控制单元状态的流动：重置门和更新门。这两个门都是可学习的参数，它们允许 GRU 选择要从上一个时间步的隐藏状态中保留多少信息，以及要从当前时间步的输入中接受多少新信息。

如图 2-6 所示。r_t和z_t分别表示重置门和更新门。重置门的作用是控制前一状态有多少信息被写入到当前的候选集$\widetilde{h_t}$上，以决定丢弃多少过去的信息。r_t越小，说明丢弃的信息越多；r_t越大，则上一时刻需要记住的信息也就越多。更新门是用于更新信息的门控单元，它控制历史状态信息能有多少被保留到当前状态中，并从候选状态中接受多少信息。GRU 网络的计算公式如（2-14）~（2-17）所示。

重置门 r_t计算公式：

$$r_t = \sigma(W_t x_t + U_t h_{t-1}) \tag{2-14}$$

更新门z_t计算公式：

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{2-15}$$

候选记忆单元 $\widetilde{h_t}$ 计算公式：

$$\widetilde{h_t} = \tanh\big(W x_t + U(r_t \circ h_{t-1})\big) \tag{2-16}$$

当前记忆单元 h_t计算公式：

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \tag{2-17}$$

2.2.5 JANET

JANET[47]（just another network）是在 LSTM 的基础上提出的新的结构，只有一个遗忘门，这使得模型参数大大减少。JANET 不仅提供了计算上的节省，而且在多个公开数据集上表现超过了标准 LSTM。JANET 的模型结构如图 2-7 所示，计算公式如下：

$$f_t = \sigma\big(W_f^h \cdot h_{t-1} + W_f^x \cdot x_t + b_f\big) \tag{2-18}$$

$$g_t = \tanh\big(W_g^h \cdot h_{t-1} + W_g^x \cdot x_t + b_g\big) \tag{2-19}$$

$$C_t = f_t \odot C_{t-1} + (1 - f_t) \odot g_t \tag{2-20}$$

$$h_t = C_t \tag{2-21}$$

遗忘门f_t是 LSTM 中最重要的结构，因此 JANET 保留了该结构。f_t决定了哪些信息应该被丢弃，因此，$1 - f_t$近似被视为i_t，减少了输入门引起的参数和计算。输出门选择记忆状态C_t中的有用信息并将其传递给隐藏状态h_t。事实上，这个任

务可以交给下一时刻的遗忘门来完成。基于这个想法，JANET 取消了输出门，合并了隐藏状态h_t和记忆状态C_t。这使得 JANET 的结构更简单，计算量更少。

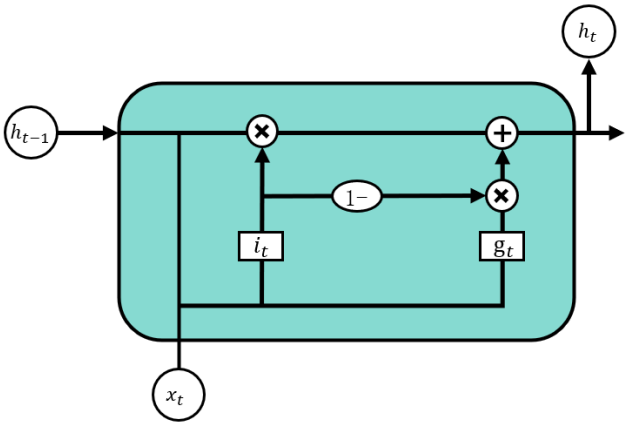

图 2-7 JANET 结构图

2.3 注意力机制

注意力机制(Attention Mechanism)是一种广泛应用于深度学习中的技术，可以使模型更加专注于与任务相关的信息。它的基本思想是在计算模型输出时，对输入的不同部分分配不同的权重，以便更好地聚焦于与任务相关的信息，从而提高模型的性能和精度，如图 2-8 所示。

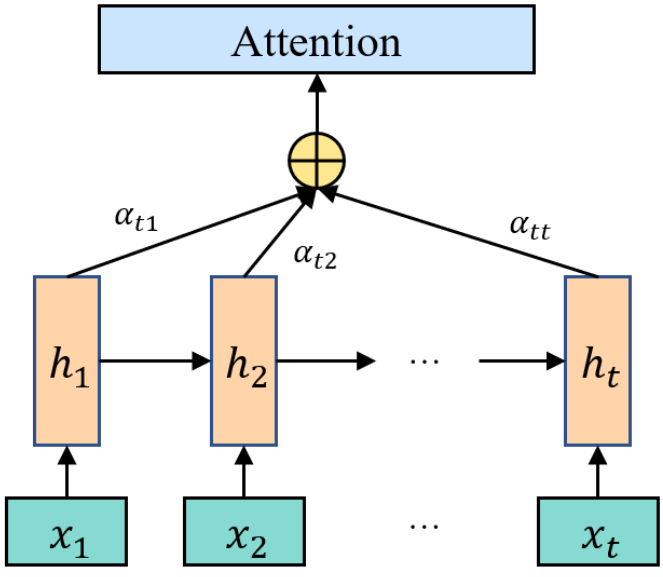

图 2-8 Attention 结构图

在计算机视觉和自然语言处理领域，注意力机制已成为一种广泛采用的技术，特别是在处理序列数据时。该机制使得模型可以在处理序列数据时，基于不同部分的重要性为它们分配不同的权重，从而使模型能够将注意力集中在最有用的部分。这种机制能够使得模型更加智能化和适应性强，因为它能够动态地调整注意力的分配方式，根据输入数据的不同特点和不同部分的重要性来优化模型的性能和精度。计算公式如下所示，其中α_{ti}即为对输入向量的权值分配。

$$\alpha_{ti} = \frac{\exp\left(similarity\left(h_i, h_j\right)\right)}{\sum_{i=1}^{t} \exp\left(similarity\left(h_i, h_j\right)\right)} \qquad (2-22)$$

2.4 三支决策基础理论

三支决策是一种决策方法，它在传统的二支决策模型基础上引入了一种延迟决策的选项。它源于粗糙集理论，是一种不确定性决策方法。在三支决策模型中，每个决策对象有三种可能的结果：接受、拒绝或延迟。

三支决策模型的引入是为了解决在现实生活中决策过程中经常遇到的信息不充分或证据不足的情况。在这种情况下，强制做出接受或拒绝决策可能会产生不必要的代价或导致严重后果。相比较于传统的二支决策，三支决策采用延迟决策策略，对于某些对象，暂时不做出决策，而是等待更多、更充分的信息或证据以便进行更为准确的判断。三支决策的实际应用非常广泛。例如，在医学领域，对于一些复杂的病例，医生可能需要进行进一步的检查和测试，以便做出最为准确的诊断和治疗方案。在金融领域，投资者可能需要等待更多的市场信息以便进行更为明智的投资决策。在工业生产中，生产商可能需要进一步的测试和验证以便确保产品的质量和安全性。

定义 2.1 给定一个决策信息系统$DS = (U, C \cup D, V, f)$，假定R是 DS上的一个等价关系，X为的U一个子集$X \epsilon U$。$\Omega = \{X, X^C\}$为X的两种状态，表示对象x是否属于目标概念。$A = \{a_P, a_B, a_N\}$表示对象x的三种决策行为，a_p表示做出接受决策的行动，即将对象X划入正域$POS(X)$ ； a_B表示做出延迟决策的行为，即将对象x划入边界域$BND(X)$ ；a_N表示做出拒绝决策的行为，即将对象

x划入负域$NEG(X)$。表 2-1 给出了对象x在两种不同状态下采取不同决策行动的损失函数：

表 2-1 不同决策行动的损失函数

	a_p	a_B	a_N
X	λ_{PP}	λ_{BP}	λ_{NP}
X^c	λ_{PN}	λ_{BN}	λ_{NN}

其中， λ_{PP} ，λ_{BP},和 λ_{NP},表示在对象 x 真实属于 X 状态情况下，做出a_P,a_B和a_N三种不同决策行为的风险代价；同理， λ_{PN} ，λ_{BN},和 λ_{NN},表示在对象 x 真实属于状态 X^C 情况 下，做出a_P,a_B和a_N三种不同决策行为的风险代价。这时，对象 x 采取 a_P,a_B和a_N三种不同决策行为的损失函数分别为：

$$R(a_P \mid [x]_R) = \lambda_{PP}\Pr(X \mid [x]_R) + \lambda_{PN}\Pr(X^c \mid [x]_R) \quad (2-23)$$

$$R(a_B \mid [x]_R) = \lambda_{BP}\Pr(X \mid [x]_R) + \lambda_{BN}\Pr(X^c \mid [x]_R) \quad (2-24)$$

$$R(a_N \mid [x]_R) = \lambda_{NP}\Pr(X \mid [x]_R) + \lambda_{NN}\Pr(X^c \mid [x]_R) \quad (2-25)$$

其$\Pr(X \mid [x]_R)$中表示 x 属于 X 状态的概率，$\Pr(X^c \mid [x]_R)$表示 x 属于 X^C状态的概率，则有$\Pr(X \mid [x]_R) + \Pr(X^c \mid [x]_R) = 1$ 。根据贝叶斯决策准则，我们需要选择期望值损失最小的决策，此会时会得到如下决策准则：

(P1)如果$R(a_P \mid [x]_R) \le R(a_B \mid [x]_R)$ 且$R(a_P \mid [x]_R) \le R(a_N \mid [x]_R)$ ，则$x \in POS(X)$；

(B1)如果$R(a_B \mid [x]_R) \le R(a_P \mid [x]_R)$ 且$R(a_B \mid [x]_R) \le R(a_N \mid [x]_R)$ ，则$x \in BND(X)$；

(N1)如果$R(a_N \mid [x]_R) \le R(a_P \mid [x]_R)$ 且$R(a_N \mid [x]_R) \le R(a_B \mid [x]_R)$ ，则$x \in NEG(X)$。

从公式(2-23)、(2-24)和(2-25)可知，以上三种决策准则只与$\Pr(X \mid [x]_R)$和损失函数有关。在三支决策模型中，将一个对象x划入不同的模糊集合会产生不同的损失。具体而言，将真实属于状态X的对象x划入正域$POS(X)$的损失最小，如果划入负域$NEG(X)$，则损失最大。划入边界域$BND(X)$的损失则介于正域$POS(X)$和负域$NEG(X)$之间。因此可以推导出$0 \le \lambda_{PP} \le \lambda_{BP} \le \lambda_{NP}$，$0 \le \lambda_{NN} \le \lambda_{BN} \le \lambda_{PN}$。这时可以将上述三条决策准则改写为：

(P2)如果$\Pr(X \mid [x]_R) \ge \alpha$且$\Pr(X \mid [x]_R) \ge \beta$时，则$x \in POS(X)$；

(B2)如果$\beta < \Pr(X \mid [x]_R) < \alpha$，则$x \in BND(X)$

(N2)如果Pr (X | $[x]_R$) $\leq \beta$且Pr (X | $[x]_R$) $\leq \gamma$时，则$x \in NEG(X)$。

其中α，β，γ的计算公式为：

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})} \quad (2-26)$$

$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})} \quad (2-27)$$

$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})} \quad (2-28)$$

由规则(B1)可知，$\beta < \alpha$，因此有$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}} < \frac{\lambda_{AP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$，再有$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}} < \frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{PN}-\lambda_{NN}} < \frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$。这时可求得$\alpha$，$\beta$，$\gamma$的关系为：$0 \leq \beta \leq \gamma \leq \alpha \leq 1$。则上述(P2)-(N2)可写为：

(P3) 如果Pr (X | $[x]_R$) $\geq \alpha$，则$x \in POS(X)$；

(B3) 如果$\beta <$ Pr (X | $[x]_R$) $< \alpha$，则$x \in BND(X)$；

(N3) 如果Pr (X | $[x]_R$) $\leq \beta$，则$x \in NEG(X)$；

事实上，如果$\beta = \alpha = 0.5$时，三支决策转变为二支决策。

2.5 本章小结

本章主要介绍了本文所涉及的相关技术和算法。其中包括 BERT 模型，以及 BERT 对句向量编码和下游分类任务的介绍。此外，还对神经网络模型、注意力机制以及三支决策理论基础进行详细描述。这些技术和算法在自然语言处理领域中应用广泛，对于本文的实验设计和实现具有重要的指导意义。

3.基于数字化互动媒体驱动的股指波动预测

3.1 问题描述

中国股票市场的稳定不仅对上市企业的未来发展和人民生活改善有着深远的影响，同时也是国家长治久安和经济持续发展的重要基石。因此，准确预测股票市场的波动趋势对于有效管控金融风险、指导投资决策以及市场监管行为具有极其重要的意义。

随着计算机技术的高速发展和互联网的普及，网络已经成为了投资者获取信息的最广泛来源。在此过程中，上证 e 互动，深交所互动易等数字化互动媒体成为了一种新型的信息传播方式。这种媒体形式基于上市公司与投资者之间的问答交互，投资者可以通过向上市公司提问来获取相关信息，而上市公司则需要对提问进行回答。这种信息交互传导机制可以影响投资者的市场参与行为，并最终在股票价格上体现出资本市场对问答信息的解读。

本文主要研究如何通过数字化互动媒体对中国股票市场进行股指涨跌预测，在本文中主要考虑以下几个问题：

（1）如何有效地量化互动媒体的信息？目前的大多数研究还是将新闻媒体或者社交媒体作为研究对象，针对新兴的数字化互动媒体的研究涉及较少。数字化互动媒体新形式的出现，对媒体信息的特征挖掘提出了新的挑战。事实上，在不同的交互模式下，即使是相似内容的互动媒体信息，对证券市场带来的影响也可能大相径庭。因此，除了对媒体信息内容的量化，我们还有必要从数字化互动媒体信息中挖掘出独特的互动特征。

（2）如何及时的提取情感特征？特征的选择对于预测结果的提升是至关重要的一步。互动平台拥有庞大的用户量，在活跃时段每分钟能产生大量的

问答文本数据。相比于内容特征，情感特征需要使用深度学习模型进行推理，时间相比于统计方法、机器学习模型较长。这可能会导致推理出结果的时间，已经迟于预测时间，从而无法使用该文本的问答情感值，失去了该文本的情感价值。

（3）如何有效地构建预测模型？金融市场的股价预测是一个具有挑战性的问题，因为金融市场的波动受多种因素的影响，如宏观经济状况、公司财务状况、市场情绪、政治环境等，这些因素的变化往往是不可预测的。此外，金融市场的行为也具有一定的不确定性和随机性，股价的变化不仅受到市场基本面因素的影响，还受到市场投资者的情绪和行为的影响，这使得股价预测变得更加困难。

本文为了解决以上问题，提出了以下方案。

（1）本文在广泛文献调研的基础上，综合考虑数字化互动媒体的特点，选取前人已证实会产生影响的特征作为显式交互特征。此外，将问答句子对使用 BERT 词向量编码，构建问答矩阵，使用 CNN 对矩阵提取特征作为隐式交互特征。以上两者结合作为互动特征。

（2）本文使用 BERT 模型对文本进行情感值判断，并在 BERT 中的每一个 Encoder 层引入三支决策，以误分类与超时决策作为代价，选取合适的阈值，使得模型能够在保证准确性的前提下，尽可能的快速输出情感分类指标。

（3）本文提出了 Bi-JANETA 预测模型。JANET[47]是在传统的 LSTM 上作了改进，其取消了输出门，合并了隐藏状态和记忆状态，并减少了输入门引起的参数和计算。这使得 JANET 的结构更简单，计算量更少，训练的时间更短。本文在 JANET 的基础上引入双向循环神经网络与注意力机制，进一步提高了模型的性能。具体来说，双向循环神经网络能够处理输入序列中的正向和反向信息，从而更好地捕捉长期依赖关系；而注意力机制则能够自适应地关注输入序列中的重要信息，有助于模型更好地理解上下文信息。这些改进使得 Bi-JANETA 模型具有更好的泛化能力和预测精度。

图 3-1 是具体研究框架图。

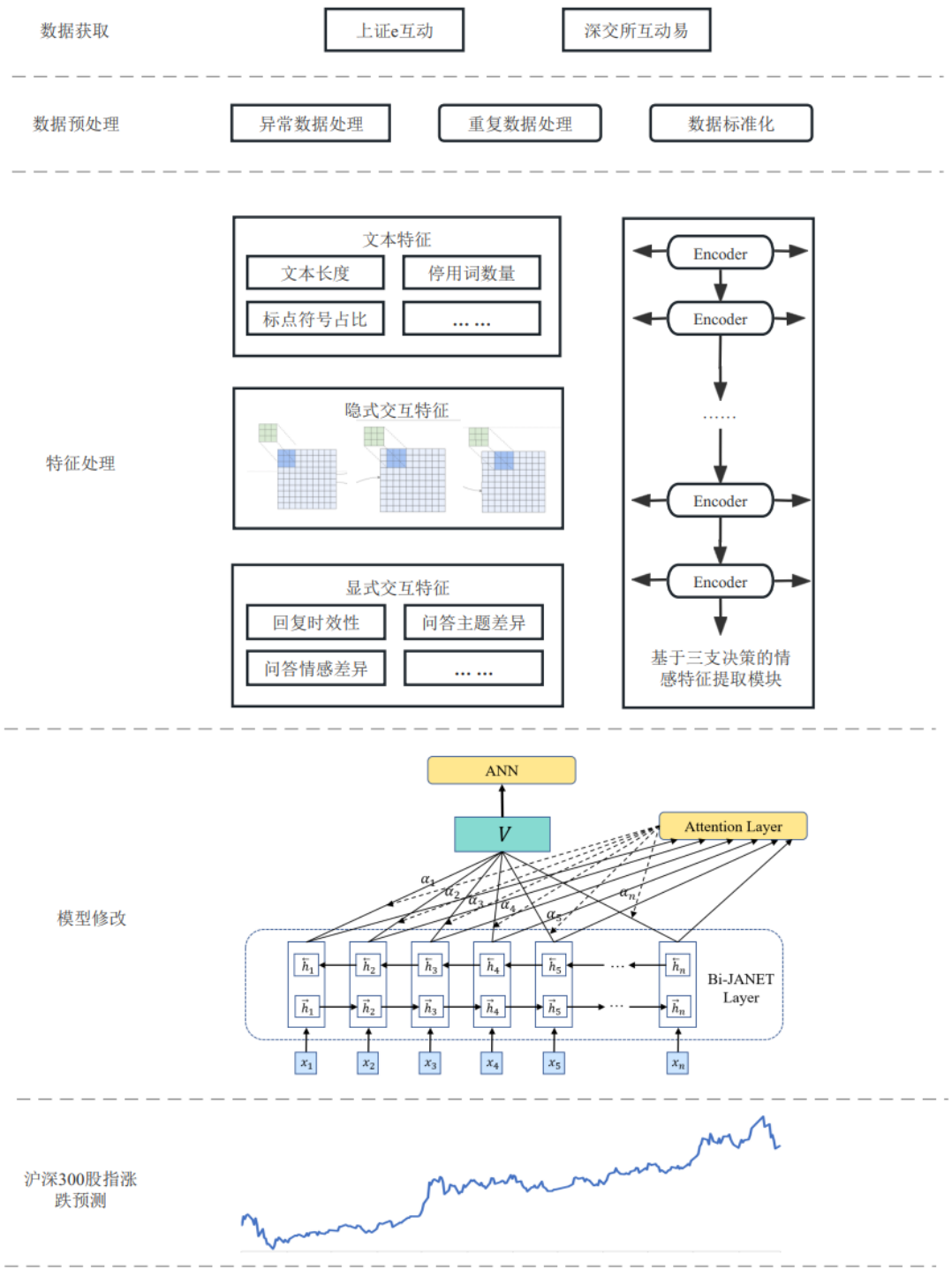
数据获取
上证e互动
深交所互动易
数据预处理
异常数据处理
重复数据处理
数据标准化
特征处理
文本特征
文本长度
停用词数量
标点符号占比
... ...
隐式交互特征
显式交互特征
回复时效性
问答主题差异
问答情感差异
... ...
Encoder
Encoder
.....
Encoder
Encoder
基于三支决策的情感特征提取模块
模型修改
ANN
V
Attention Layer
Bi-JANET Layer
沪深300股指涨跌预测

图 3-1 研究框架图

3.2 问答文本数据获取及描述

本文选取上证“e 互动”[1]投资者互动平台、深交所“互动易”[2]两大主流互动平台的问答内容作为互动文本数据的来源。一是因为两者皆是官方建立的互动问答平台，且成立时间较长，用户数量庞大。二是庞大的用户量给平台带来了较高的日活跃度，每天都有大量的问答文本数据。三是相比于其他社交和新闻媒体，互动平台中的用户主要是投资者或上市公司，所讨论的事情与股票市场息息相关，可以排除一定的网络噪声。

获取互动媒体问答文本数据的主要步骤有以下 3 步：

第 1 步为文本数据收集。本文使用 Python 语言，基于 Scrapy 框架自定义网络爬虫，爬取上证“e 互动”与深交所“互动易”平台从 2020 年 3 月 1 日到 2021 年 3 月 1 日的文本数据，每条数据包括了上市公司代码、提问内容、提问时间、回复内容、回复时间等信息，共计 467,882 条数据。部分数据如图 3-2 所示。

002583	公司口罩日产量是多少？	2020-03-06 14:47:41	您好，鉴于全球疫情情况和防控需	2020-03-09 09:43:23
002443	近几年贵司的股价一路向低，虽然说中国上市公	2020-03-05 15:12:58	也许是因为公司踏实生产，不蹭热	2020-03-09 09:43:27
300756	一季度游戏收入会不会增加	2020-03-08 19:41:39	您好！公司是一家专业从事游乐设	2020-03-09 09:43:59
002177	公司有没有生产口罩的打算，减少atm生产线，扌	2020-03-06 14:37:08	感谢您的建议，谢谢关注！	2020-03-09 09:45:25
300767	请问建设工程抗震管理条例公布后，公司的产能	2020-03-06 19:25:24	尊敬的投资者您好。该条例目前已	2020-03-09 09:45:32
002925	请问公司的工业互联网和物联网发展情况怎么样	2020-03-05 09:13:53	您好。（1）公司的工业互联网包括	2020-03-09 09:45:35
002950	公司今年如何利用产能扩大收入和利润，销售方	2020-02-20 14:08:58	尊敬的投资者您好！目前，公司正	2020-03-09 09:46:25
300664	请问 艾棣维欣 和 Inovio 的合作协议是否已签署？	2020-03-07 16:50:57	你好，公司正持续跟踪投资艾棣维	2020-03-09 09:46:29
002190	你好，请问公司目前复工复产率有多少？今年一	2020-03-05 10:30:21	您好，公司已全面复工，一季度数	2020-03-09 09:46:33
300079	请问公司，2019年半年报中，关于数字电视类的	2020-03-05 10:10:26	您好，公司在2019年随政策及市场	2020-03-09 09:47:10
002925	请问董秘公司有涉足半导体行业吗？芯片业务有	2020-03-05 16:36:25	您好。（1）公司以自主创新的UDN	2020-03-09 09:47:23
300079	公司这么长时间了，为什么还是郑董事长身兼总	2020-03-04 22:28:26	您好，我们会积极关注您提到的问	2020-03-09 09:47:26
300079	郑董，您好！公司在2017、2018两个年度报告都	2020-03-04 10:27:35	您好，相关工作持续推进当中，如	2020-03-09 09:47:50
002957	董秘你好，目前口罩机订单量是否可以披露？对	2020-03-02 11:07:57	您好！公司生产口罩机设备主要是	2020-03-09 09:47:56
300079	目前股价低迷，公司号称拥有大量现金流却不积	2020-03-04 10:24:41	您好，我们会按照规则披露回购进	2020-03-09 09:48:11
300079	贵公司在3月2日回购了29.89万，回购时间只剩下	2020-03-03 19:52:17	您好，我们会按照规则披露回购进	2020-03-09 09:49:07
300079	郑董事长您好，看了下您在投资者互动平台上关	2020-03-03 15:50:32	您好，商业细节未触及披露义务，	2020-03-09 09:49:57
300079	1、请问疫情对于公司一季度业绩有多大影响？2	2020-03-02 15:51:51	您好，1、本公司正在努力降低疫情	2020-03-09 09:50:37

图 3-2 部分问答数据

第 2 步为文本预处理。文本预处理是自然语言处理的一个重要步骤，能够对原始数据进行清洗和筛选，提高后续处理的效率和准确性。在本文中的文本预处理主要包括以下几个方面：

（1）清洗无效符号和字符乱码数据：在文本数据中，可能存在一些无效符号和字符乱码数据，需要将其清洗掉，以避免对后续处理产生影响。

[1] http://sns.sseinfo.com/
[2] http://irm.cninfo.com.cn/

（2）删除相同内容、时间的样本：在互动平台中，可能会存在一些重复的讨论，这些重复的讨论没有实际意义，需要将其删除。

（3）剔除字数过长、过短样本：在文本数据中，一些字数过长或过短的样本可能会影响后续的分析和建模，需要将其剔除。

（4）剔除空白样本：一些只包含空白符号的样本对于分析和建模没有意义，需要将其剔除。

（5）删除广告贴：互动平台中可能存在一些广告贴，这些贴子对于分析和建模没有实际意义，需要将其删除。

通过对原始数据进行上述预处理，可以使得后续处理更加准确和高效，提高模型的性能和准确度。经数据处理后最终得到有效数据 461,371 条，日均数据约 1,264 条，月均数据约 35,490 条。图 3-3 为每个月的数据量。从图中可以看出，在 2021 年 2 月数据量有所下降，这可能是受到了春节假期的影响。但是总体上一年以来的文本数据在持续上升，这表明了互动媒体的活跃度与用户量正逐步上升。

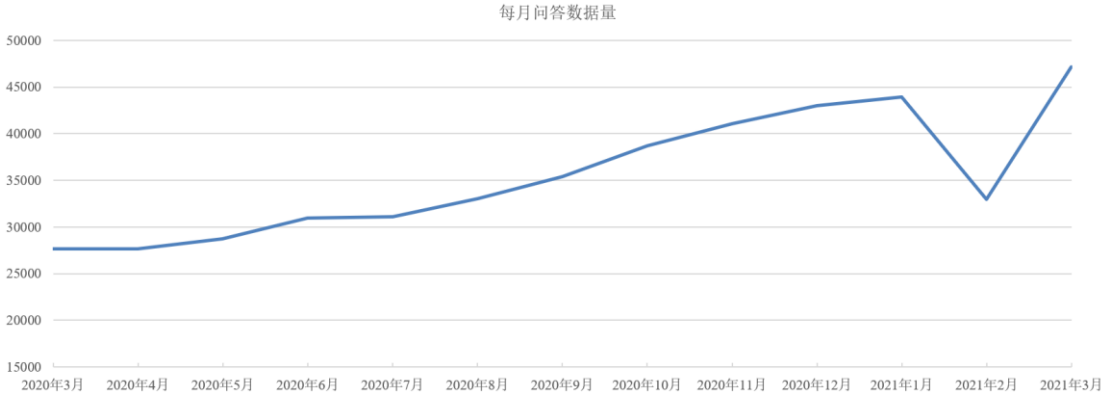

图 3-3 问答数据量（月）

第 3 步为文本的特征处理。本文选取的特征为内容特征、显式交互特征、隐式交互特征、情感特征。其中内容特征包括文本长度、文本停用词数量、字符数等。显式交互特征包括问答句子对中文本长度比值、回复时效性等。隐式交互特征为使用 CNN 提取基于问答文本矩阵特征。情感特征为文本中积极或消极的情感值。具体步骤在下一部分做详细解释。