Advances in Applied Mathematics 应用数学进展, 2025, 14(11), 117-122
Published Online November 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.1411467

Hans汉斯

基于BiLSTM的证券市场情绪感知方法

李　晴*，申　晨#，高瑞娟，付如意

河北金融学院河北省金融科技应用重点实验室，河北　保定

收稿日期：2025年10月7日；录用日期：2025年11月1日；发布日期：2025年11月7日

摘　要

在证券市场中，市场情绪对投资决策的影响日益显著。基于BiLSTM模型，设计了一种证券市场情绪感知方法，通过数据爬取、数据预处理、词向量构建、情绪分类等步骤，实现了情绪感知。实验表明，该方法能有效捕捉投资者情绪倾向，验证了BiLSTM模型在证券市场情绪感知中的有效性，为投资者提供决策参考。

关键词

BiLSTM，情绪感知，证券市场

BiLSTM-Based Sentiment Analysis Approach for Securities Market

Qing Li*, Chen Shen#, Ruijuan Gao, Ruyi Fu

Hebei Key Laboratory of Financial Technology Application, Hebei Finance University, Baoding Hebei

Received: October 7, 2025; accepted: November 1, 2025; published: November 7, 2025

Abstract

In the securities market, market sentiment has an increasingly significant impact on investment decisions. This paper proposes a BiLSTM-based sentiment analysis method for the securities market, which achieves sentiment perception through key steps including data crawling, data preprocessing, word embedding generation, and sentiment classification. Experimental results demonstrate that the method can effectively capture investor sentiment tendencies, validating the effectiveness of the BiLSTM model in securities market sentiment analysis and providing decision support for investors.

*第一作者。
#通讯作者。

文章引用: 李晴, 申晨, 高瑞娟, 付如意. 基于 BiLSTM 的证券市场情绪感知方法[J]. 应用数学进展, 2025, 14(11): 117-122. DOI: 10.12677/aam.2025.1411467

Keywords

BiLSTM, Sentiment Analysis, Securities Market

1. 引言

1990 年，上海证券交易所和深圳证券交易所相继成立，标志着新中国集中交易的证券市场正式诞生。三十多年来，伴随着经济体制改革和市场经济发展，我国证券市场制度不断健全、体系不断完善、规模不断扩大，已经成为我国经济体系的重要组成部分。截至 2024 年 12 月，中国证券市场投资者数量已突破 2.2 亿户，其中自然人投资者占比超过 99%。庞大的个人投资者群体其情绪波动极易形成“羊群效应”，对市场稳定构成潜在挑战。对投资者而言，情绪感知方法能够量化市场中的非理性情绪，帮助其识别群体心理偏差，优化投资决策并规避风险[1]-[3]；对监管机构而言，通过实时监测网络舆情中的情绪倾向，可提前预警市场异常波动，提升监管精准度，维护金融稳定[4]-[6]；因此具有重要的理论和现实意义。

本文首先通过网络爬虫技术采集东方财富网股吧中的文本信息作为数据来源，并对数据进行去重、标注、分词、去停用词、生成词向量等操作，然后使用以上数据，选取 BiLSTM 模型进行分类训练和测试，最后通过准确率、精准率、召回率和 F1 值、交叉熵损失函数等指标对模型性能进行评估。

2. 数据获取与处理

2.1. 数据获取

本文选择东方财富网股吧中的上证指数吧(zssh000001)作为数据来源，采用 Python 语言爬取 2023~2024 年帖子中相关信息，爬取过程总共分为五个步骤[7]。

第一步对东方财富网股吧网页进行分析，使用 lxml 库解析 html，提取阅读量、帖子内容、作者、发帖时间四个元素，主要实现代码如下：

```
read=selector.xpath('//div[contains(@class,"articlehnormal_post")]//span[@class="l1 a1"]//text()')
title=selector.xpath('//div[contains(@class,"articlehnormal_post")]/span[@class="l3 a3"]/a')
author=selector.xpath('//div[contains(@class,"articlehnormal_post")]//span[@class="l4 a4"]//text()')
date=selector.xpath('//div[contains(@class,"articlehnormal_post")]//span[@class="l5 a5"]//text()')
```

第二步对于要爬取的页面信息，确定 URL 集，再根据要爬取页面，对 URL 进行数据分析。

第三步创建虚拟用户代理并向服务器发送请求，下载所需内容。

第四步解析和下载阅读量、帖子内容、作者、发帖时间，共四项数据。

第五步以 CSV 格式存储爬取到的数据。

2.2. 数据描述

爬取到的数据主要包括阅读量(read)、帖子内容(title)、作者(author)、发帖时间(date)共四个字段。部分数据如表 1 所示。

Table 1. Partial sample data
表 1. 部分样本数据

阅读量(read)	帖子内容(title)	作者(author)	发帖时间(date)
593	后市注意 3341 缺口处压力	乐奇东尼	2023/3/1 18:22:00
767	2.27 一周看法!3200 左右慢慢买!	专业股民 15888	2023/3/1 17:54:00
125	刚准备退市呢，你就涨了，有本事一直跌啊	逆人性的弱点	2023/2/28 19:45:00

2.3. 数据预处理

数据预处理部分的主要工作包括：去重处理，数据标注，分词，去停用词，文本向量化，以方便后续计算。数据预处理的基本框架如图 1 所示。

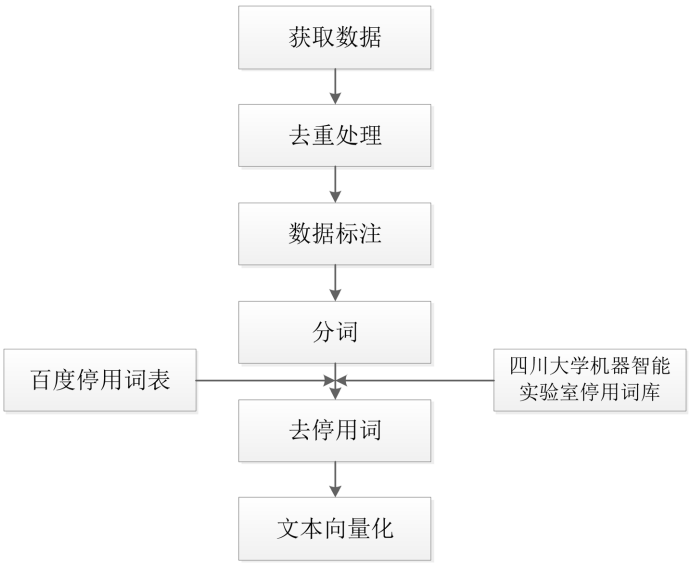

Figure 1. Data preprocessing pipeline
图 1. 数据预处理框架图

首先，对爬取的股吧文本数据进行去重处理。针对数据中存在的大量重复值，采用 pandas 库的 drop_duplicates()函数，以 title 列为基准进行去重。同时，利用 isin()函数识别并处理包含图片链接、转发链接等非文本内容的记录，保留首次出现的有效数据。此外，为提升数据质量，还通过正则表达式 pattern = r"^\d+$"清除了纯数字等无意义文本。最终得到约 7200 条数据。

其次，对文本数据进行手动标注。消极情感文本数据标记为 0，积极情感文本数据标记为 1，中性情感文本数据标记为 2。

再次，分词和去停用词。用户评论中常包含“@”、“#”、“*”等非语义符号，这些符号会增加数据维度，干扰模型训练，在数据预处理阶段需将其清除。本研究首先采用 jieba 库的精确分词模式对文本进行切分，随后结合百度停用词表与四川大学机器智能实验室停用词库，统一过滤掉特殊符号、停用词及语气助词等冗余信息。处理后的数据更加精确，有利于后续模型的文本分析。

最后，进行词向量训练。考虑到 BERT 模型对输入句长的限制(通常为 512 字符)，首先利用 Python 对预处理后的数据进行句长分析。分析结果显示，数据集中所有评论的句长均符合该要求。在此基础上，将数据输入 BERT 模型进行词向量训练。该过程利用 BERT 自带的中文词典，将每个句子转换为一个由

高维词向量组成的序列(Sequence of Word Vectors)。这个序列中的每个向量都动态地融合了其上下文的语义信息，为后续的序列模型(BiLSTM)提供了高质量的输入。

3. 基于 BiLSTM 的情绪感知模型构建

3.1. 模型选择与分析

为精准感知投资者情绪，本研究构建了融合 BERT 词向量与 BiLSTM 序列模型的深度学习框架。在特征提取层面，相较于传统机器学习依赖 TF-IDF 等静态特征，本研究采用的 BERT 词向量能够动态融合上下文语义，更深刻地捕捉金融文本的复杂内涵。在模型结构层面，BiLSTM 的双向架构使其能同时捕获前向与后向的文本依赖，相较于单向 LSTM，能更准确地解析股评中常见的转折、因果等复杂句式，从而提升对情感色彩的识别精度。此外，相较于直接微调 BERT 模型，本研究的“BERT + BiLSTM”混合策略在保证高性能的同时，降低了对大规模标注数据的依赖与计算成本，更具备实际应用中的灵活性与效率。

3.2. 数据准备与划分

在模型训练前，首先对经过预处理的文本数据集进行科学划分。利用 scikit-learn 库中的 train_test_split 函数，将数据集按 8:2 的比例随机分割为训练集与测试集。其中，80%的数据用于训练模型，使其学习文本特征与情绪标签之间的映射关系；剩余 20%的数据作为独立的测试集，用于在训练完成后评估模型的泛化能力和性能表现。为确保实验结果的可复现性，通过设置 random_state=123 固定了随机种子，确保每次运行时数据划分方式完全一致。具体实现代码如下：

```
xtrain,xtest,ytrain,ytest=train_test_split(texts,labels,test_size=0.2,random_state=123)
```

3.3. 模型架构设计与参数配置

模型的关键参数定义如下：

hidden_size (隐藏层维度)：设定为 768。此参数的选择与 BERT 模型生成的词向量维度保持一致，以确保数据流在模型各层之间顺畅传递，并充分利用预训练模型提取的深层语义特征。

num_layers (网络层数)：设定为 1。考虑到本研究的数据集规模以及模型复杂度，选择单层 BiLSTM 结构，以保证模型具备足够学习能力的同时，有效控制参数数量，降低过拟合的风险。

epochs (训练轮次)：设定为 32。设置足够的训练轮次旨在使模型具备充足的学习时间，使其能够逐步优化参数，最终收敛到一个性能较优的状态。

模型构建的核心实现代码如下：

```
self.bert = BertModel.from_pretrained('bert-base-chinese')
self.lstm = nn.LSTM(input_size=model_config.hidden_size,
                    hidden_size=model_config.hidden_size // 2,
                    num_layers=1,
                    bidirectional=True, #  双向
                    batch_first=True)
self.classifier = nn.Linear(model_config.hidden_size, num_classes)
```

3.4. 模型优化策略

本研究采用了 AdamW 优化器，通过迭代计算梯度并更新模型参数，以最小化预设的损失函数。

AdamW 是 Adam 算法的改进版本，它引入了权重衰减机制，能更有效地防止模型过拟合。此处将学习率设定为 0.005，旨在平衡收敛速度与训练稳定性。Adam 优化器的主要优点包括：实现简便、计算效率高、对内存需求较低，并且其自适应学习率的特性使其对梯度的伸缩变化不敏感，非常适合处理复杂的非凸优化问题。此外，还配置了一个带预热机制的学习率调度器，以在训练初期稳定学习过程。其核心实现代码如下：

```
optimizer= torch.optim.AdamW(optimizer_grouped_parameters, lr=config.lr)
scheduler = transformers.get_linear_schedule_with_warmup(optimizer,
parser.add_argument("--lr", type=float, default=0.005)
```

4. 模型评估

采用交叉熵损失函数(Cross-Entropy Loss)、精准率(Precision)、召回率(Recall)、F1 值(F1-score)在每个训练轮次(Epoch)的变化，全面、深入评估前文构建的证券市场情绪分析模型训练过程中的性能变化。

随着训练的进行，模型在训练集和测试集上的精准率与召回率均波动上升，并逐渐趋于稳定，说明模型不仅预测结果的准确性(精准率)在提高，其识别出所有相关样本的能力(召回率)也在增强；F1 值同样呈现出持续上升并最终收敛的态势，最终稳定在 80%以上，说明模型在查准(Precision)与查全(Recall)之间取得了良好的平衡，具备了可靠的分类能力。

模型在测试集上的详细运行结果如表 2 所示。

Table 2. Model performance report
表 2. 模型性能报告

	Precision	Recall	F1-score	Support
0	0.8793	0.8919	0.8856	629
1	0.8517	0.8207	0.8359	329
2	0.8165	0.8216	0.8190	482
accuracy	-	-	0.8521	1440
macro avg	0.8492	0.8447	0.8468	1440
weighted avg	0.8520	0.8521	0.8519	1440

综合上述评估指标，本研究验证了基于 BiLSTM 的情绪感知模型在证券市场情绪分类任务中的有效性。模型总体准确率为 85.21%，加权平均 F1 值为 0.8519，说明模型对股民情绪的整体判别能力与鲁棒性较强；宏平均 F1 值为 0.8468，表明在类别不平衡条件下，模型在识别不同情绪类别时仍具有均衡性和有效性。

模型对“消极”情绪的判别能力相对突出，可能得益于该类别样本相对充分，也印证了模型对关键负面信号的敏感性。同时，模型在识别“积极”和“中性”情绪时也表现出稳健的性能，避免了因样本量差异导致某些类别识别精度显著下降的问题。

5. 进一步分析

在此基础上，进一步尝试采用该模型构建日度市场情绪指数，并检验该指数与上证指数收益率、波动率或交易量之间是否存在领先或同步关系。对于日度市场情绪指数，一个较常用的方法是将市场情绪指数定义为：(积极帖数 − 消极帖数)/总帖数。初步分析表明，该指数与上证指数收益率、波动率或交易

量之间并未表现出显著的统计相关性。可能的原因主要包括两个层面，一是前文所提出的方法对评论中一些反讽和复杂语气的理解正确率还不高(如“漂亮，又创新低！”)，市场情绪指数的构造方法比较简单，未考虑帖子的阅读量、影响力以及中性帖子的信息价值，且评论的情绪无法代表市场中很少发声的“沉默的大多数”，最终导致情绪指数可能并未完整刻画市场情绪；二是证券市场中情绪与价格波动常呈现一种内生性的反馈循环，情绪指数在某些情况下可能是一个与市场同步甚至滞后的指标，其功能更多在于确认已发生的市场行情，而非预测未来走势。

尽管如此，该模型仍然为感知复杂金融文本情绪提供了相对高效、可靠的手段，在辅助投资者优化决策、防范风险，以及支持监管机构快速评估特定政策或事件的影响，进行市场风险预警等方面具有一定的应用前景[8]。

基金项目

2025 年度河北省金融科技应用重点实验室课题：融合时序特征与监管规则量化的证券市场异常交易监测研究(课题编号：2025010)。

参考文献

[1] 王娜, 贺毅岳, 刘磊. 股市投资者情绪指数构建及其有效性研究——基于东方财富股吧帖文的情感分析[J]. 价格理论与实践, 2022(11): 146-151.

[2] 周小波. 基于情感分析的中国股市短期价格预测研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2020.

[3] 卢雪姣. 网络舆情与投资者情绪的相关及溢出效应研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2023.

[4] 王超, 何建敏, 姚鸿. 基于社会网络的情绪扩散与股价波动风险研究[J]. 管理评论, 2022, 34(12): 16-25.

[5] 薛冰清. 基于文本挖掘法的中国股票市场收益与投资者情绪的关系分析[J]. 中国管理信息化, 2024, 27(20): 118-121.

[6] 籍琪. 投资者情绪对股票收益率的影响——基于文本挖掘技术的实证研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2024.

[7] 林维维. 基于投资者情绪的金融行业系统性风险测度研究[D]: [硕士学位论文]. 青岛: 青岛大学, 2022.

[8] 薛凯阳. 网络舆情、投资者情绪与股票收益率[D]: [硕士学位论文]. 北京: 商务部国际贸易经济合作研究院, 2025.