基于主题增强协同注意力 BERT 的股市多主体文本情感识别研究

赵宇鹏[1] 王文心[2]

1. 同济大学 上海 普陀 200042

1. 上海市民办平和学校 上海 普陀 200042

摘要：2023 年中央金融工作会议提出推动金融高质量发展的战略目标，在此背景下，本文聚焦 A 股市场参与者，探究管理者、分析师与投资者三类主体间的文本情感异同。并针对现有研究在跨主体情感统一测度不足的问题，本研究创新性地采用了融合主题信息的情感预测模型，基于沪深 300 成分股的 50 只股票样本对文本情感进行了识别，结果显示基于主题信息的文本情感预测模型能够有效的提升模型的预测效果，为金融领域文本情感预测提供新的思路。

关键词：股市参与者，情感识别，TescaBert 模型

DOI：10.63887/jfem.2025.1.1.18

引言

研究发现，从方法上，当前自然语言处理主要采用基于情感词典和机器学习两类方法。基于情感词典的方法通过匹配预定义词典中的情感词进行极性判断，在中文领域已有大连理工大学情感词汇本体库、HowNet 等多款通用词典。机器学习方法（如 SVM、朴素贝叶斯）在早期研究中取得较好效果（Kim 和 Kim，2014），而深度学习方法，特别是基于 Transformer 的预训练模型 BERT（Devlin，2018）及其变体（林杰，2020）在金融情感分析任务中展现出显著优势，准确率最高可达 97.35%（Li M，2021）。在应用领域中，当前文本情感对金融市场的影响研究对主要集中在单主题的情感识别上，鲜有学者对多主体情感识别任务进行关注，因此本文采用基于主题信息的文本情感预测模型，对股市参与者的情感进行预测，并检验模型效果。

一、主题增强情感协同注意力 BERT 模型

（一）BERT 模型

BERT 是由 Google 团队于 2018 年提出的基于 Transformer 架构的预训练语言模型。该模型通过大规模无监督预训练学习文本的深层语义表征，再通过微调适配下游任务，在机器翻译、情感分析等 NLP 任务中展现出卓越性能。相较于传统时序模型，BERT 支持并行计算且通过堆叠编码层增强表征能力。本文采用的 Topic enhanced sentiment co-attention BERT 模型即基于该架构，通过融入主题信息进一步优化了金融文本的情感分析性能[1]。

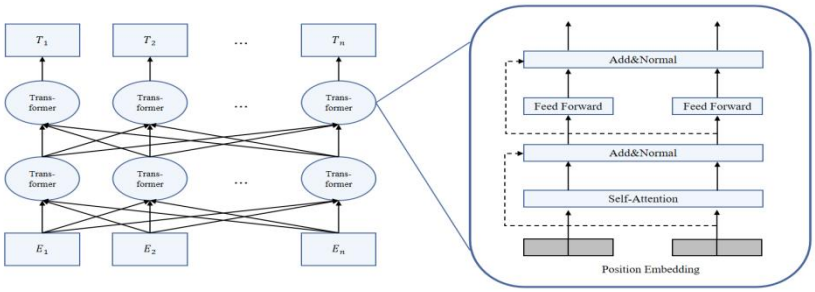

图 1 BERT 模型结构

（二）股市参与者文本特征

中国金融市场参与者可分为上市公司、证券分析师和大众投资者三类。上市公司年报采用规范书面语言，遵循"管理层讨论与分析"框架，常弱化负面信息；分析师研报专业性强，多用"估值中枢"等术语，但存在乐观偏差，70%以上为"买入"评级；而投资者股吧评论则情绪化、碎片化，充斥网络用语，时效性强但可信度低[2]。

表 1 股市参与者文本特征差异

对比维度	公司年报	分析师研报	股吧文本
语言复杂度	专业术语密集	行业黑话主导	俚语网络用语
情感极性	中性偏积极	谨慎乐观	消极为主
信息可信度	经审计但存在美化倾向	存在利益冲突潜在偏差	谣言与事实混杂
核心功能	合规披露+战略叙事	价值发现+预期引导	情绪共振+短期博弈
修辞策略	被动语态+数据堆砌	条件句+比较级论证	夸张比喻+群体暗示

三类文本在语言风格、信息含量和情感表达上的本质差异，使得传统的情感分析方法难以准确识别其真实情感倾向。特别是在金融领域，同一表述在不同语境下可能传达完全不同的情感含义。例如分析师报告中的标准风险提示语句，表面看似负面，实则为合规性表述；而年报中看似中性的表述，则可能隐藏着管理层的信息操纵意图。这种语境依赖性给情感分析带来了巨大挑战，亟需能够识别文本主题和语境的创新分析方法。

（三）股市参与者文本情感预测模型构建

如上文所述，金融文本的情感表达往往随语境和主题变化，传统的情感分类方法容易产生误判。为此，本文采用融合主题信息的 BERT 模型（TescaBERT）(Wang S,2023)，通过共同注意力机制同步学习文本主题特征和情感特征。该模型创新性地将主题表示与上下文表征输入共同注意力模块，并采用主题分类（TC）和情感预测（SA）双任务微调机制，在共享底层参数的同时实现主题知识与情感特征的协同优化，显著提升了金融领域特定语境下的情感分析准确率[3]。

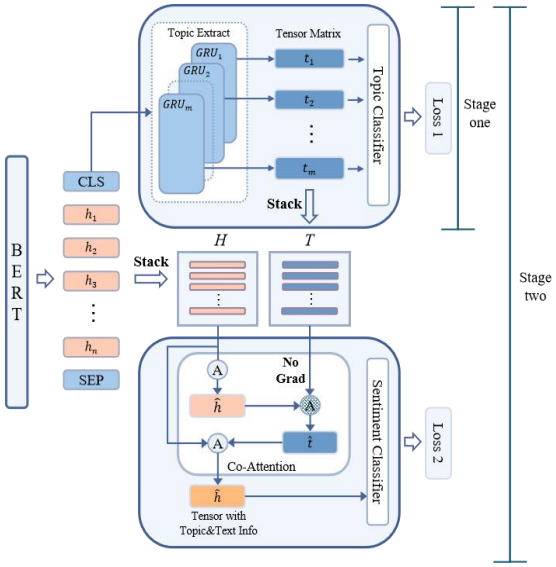

图 2 TescaBERT 模型结构

（1）第一阶段训练：主题提取(TC)

第一阶段是基于 BERT 模型以及 GRU 模型的主题分类任务。首先通过 BERT 模型进行上下文信息编码模块，将输入的文本转化为包含<CLS>标记以及<SEP>标记的张量。其次基于 GRU 模型进行主题分类训练，在模型中共设置了 M 个 GRU 模型（其中 M 为主题数量）。将文本的主题信息h_{cls}输入到 m 个 GRU 模型中，得到 m 个主题的概率分布。并用相应的主题标签计算损失用于反向传播。

（2）第二阶段训练：情感分析(SA)

第二阶段是基于 co-attention 机制的情感分析任务模块，模型采用交替共同注意机制来生成情感表征。co-attention 机制如下：

（1）对初始化 X=H，g=0，让模型训练的注意力集中在输出句子的上下文表示。

（2）令 X=T，g=$\hat{h}$。语境表征指导话题的注意力生成，该步骤生成主题注意力的表示。模型主要关注句子中主要讨论的话题，而忽略无关的话题。

（3）令 X=H，g=$\hat{t}$。主题表征指导单词的注意力生成，该步骤输出新的文本注意力表示，模型在该阶段更加关注句子中与主题相关的内容，其中包含与主题相关的情感信息。

$$M = \tanh(W_x X + (W_g g)^T) \quad (3.1)$$

$$a^x = \mathrm{softmax}(w_{hx}^T M) \quad (3.2)$$

$$\hat{x} = \sum_i a_i^x x_i \quad (3.3)$$

与此同时，为尽可能避免在第二阶段训练中引入过量的噪音，模型使用 Topic Classfier 的预测结果作为主题注意力的掩码。对于话题 i，如果预测值$\hat{y}_i^t$=0（不涉及话题 i），则话题 i 的注意力值将被掩码，使得在 i 处a_i^t=0。否则，原始注意力值将保持不变。

二、基于主题的情感预测模型构建

（一）数据提取与预处理

研究样本选取 2018-2023 年间持续入选的 50 只成分股（沪市 25 只，深市 25 只），包括贵州茅台、伊利股份等优质标的，确保数据均衡性和连续性。

本文采用爬虫从东方财富股吧、东方财富研报网、巨潮信息网以及国泰安数据库中获取了 2017 年-2023 年间标的股票对应的研报、年报以及股评。以句为单位对研报、年报进行拆分。剔除股评中重复、转发以及广告等文本，并对表情进行转化。经数据预处理后，共获取 492,699 条研报文本、26,628 条年报文本以及

780 余万条投资者评论数据。按照 6:3:1 的比例从三类文本中各抽取 15,000 条样本进行人工标注。标注采用三分类体系（2=积极，1=中性，0=负面），共获得 10,831 条有效标注数据，其中投资者评论 5,909 条、分析师报告 3,006 条、年报文本 1,916 条[4]。

表 2 文本描述性统计

	文本数量	文本长度	正向文本数量	中性文本数量	负向文本数量	合计
投资者评论	5909	19.84	1072	554	4283	5909
分析师报告	3006	62.63	1757	835	414	3006
年报	1916	150.42	1078	707	131	1916
合计	10831		3907	2069	4828	

数据揭示三类文本存在显著特征差异：（1）文本长度维度，年报文本最为详实，分析师报告次之，投资者评论最为简短；（2）情感分布维度，投资者评论呈现显著负面倾向，分析师报告保持谨慎乐观，年报则体现显著正向偏差。这种结构化差异源于各自的信息披露机制与传播特征，为构建领域自适应的情感分析模型提供了关键特征依据。

（二）模型效果

本文采用分类问题常用的准确率（Accuracy）、精准率（Precision）、召回率（Recall）和 F1-Score 作为评估结果的指标[5]。并选用支持向量机、随机森林等模型作为本文的 BaseLine 模型。从实验结果来看，TescaBERT 相比于其他的 Baseline 模型，在分类结果上有显著的提升。相比于经典的机器学习分类算法，其准确率以及 F1 值都提升了将近 20 个 BP。因此可以认为 TescaBERT 模型能够有效地提升情感分类任务的准确度。

表 3 模型结果对比

	Accuracy	Precision	Recall	F1_score
支持向量机	0.6249	0.4813	0.6249	0.5438
随机森林	0.6123	0.4411	0.6123	0.5129
BiLSTM	0.6713	0.7344	0.7200	0.7272
BERT	0.7178	0.7619	0.7929	0.7771
TescaBERT	0.8025	0.8026	0.8125	0.8075

（三）稳健型检验

为验证模型泛化性，本研究采用谭松波教授构建的 ChnSentiCorp_htl 标准数据集（正评 5,322 条，负评 2,444 条）进行测试。实验结果表明，TescaBERT 在基准模型中表现最优，证实了融合主题信息的情感预测模型在准确率上的显著优势，凸显其对本研究任务的适用性。

表 4 稳健型结果对比

	Accuracy	Precision	Recall	F1_score
支持向量机	0.7121	0.7034	0.7121	0.6303
随机森林	0.7115	0.7032	0.7115	0.6299
BiLSTM	0.7753	0.7750	0.8401	0.8062
BERT	0.8840	0.9157	0.8840	0.8842
TescaBERT	0.9214	0.9223	0.9214	0.9218

三、研究结论

本文基于 TescaBERT 模型，针对 A 股市场参与者（管理者、分析师与投资者）的多主题文本情感识别问题展开研究，提出了一种融

合主题信息的情感预测模型。通过构建主题增强情感协同注意力机制（TescaBERT），模型在金融领域文本的情感分析任务中表现出显著优势。实验结果表明，TescaBERT 在沪深 300 成分股的三类文本数据集上取得了最优性能（F1-score 达 0.8075）。进一步通过 ChnSentiCorp_htl 标准数据集的稳健性检验，验证了模型在跨领域任务中的泛化能力，表明主题信息的引入能有效解决金融文本中因语境差异导致的情感误判问题。

参考文献

[1]Kim S ,Kim D .Investor sentiment from internet message postings and the predictability of stock returns[J].Journal of Economic Behavior and Organization,2014,70-79.

[2]Devlin J, Chang M, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805,2018

[3]林杰,江晨曦.基于 BERT 模型的投资者情绪指数建模及与价格关系分析[J].上海管理科学,2020,42(4):75-80.

[4]Li M,Li W,Wang F,et al.Applying BERT to analyze investor sentiment in stock market[J].Neural Computing and Applications,2021:4663-4676.

[5]Wang S,Zhou G,Lu J,et al.Topic enhanced sentiment co-attention BERT[J].Journal of Intelligent Information Systems,2023,60(1)12-17.