

基于 FinBERT-RCNN 的金融文本情感分析模型

刘一辰

(中国人民大学信息学院, 北京 100872)

摘要: 投资者的情绪变化对我国的股市波动影响较大, 金融文本包含了投资者的情绪以及对相关事件的态度, 提供了一种了解股票市场变化的渠道。近年来, 自然语言处理已广泛应用于金融领域, 对金融文本进行情感分析可以得到丰富的投资价值和监管参考价值。然而由于金融词汇的专业性和特殊性, 现有的通用情感分析模型不适用于金融领域情感分析任务, 且单一模型提取的特征较为片面, 模型准确性有待提高。为解决这些问题, 本文提出一个 FinBERT-RCNN 模型, 利用在金融语料库预训练的 FinBERT 模型进行词嵌入处理, 提取语义特征, 将提取到的特征引入 RCNN 模型进一步挖掘上下文关键特征, 提升了模型的鲁棒性和泛化性。实验结果表明, 在金融领域数据集上, 提出的模型优于其他情感分析模型, 准确度提高了 6%~47%。

关键词: 金融文本; 情感分析; FinBERT; 循环卷积神经网络

(注: 相关代码已在 GitHub 开源, 项目链接为 <https://github.com/lyc289/FinBERT-RCNN>)

1 引言

我国的股票交易市场整体呈现出非有效性, 舆论和政策对股票市场有较大影响。根据行为金融学理论, 投资者并不都是理性投资者, 在投资过程中会出现受他人情绪影响而改变投资倾向的行为。金融领域的文本数据包含对相关事件的情绪, 对金融文本进行情感分析有助于理解投资者的态度, 并影响投资决策和市场走势。自然语言处理的情感分析技术对带有情感色彩的主观性文本进行分析, 可以从文本中提取语义情感信息。近年来, 自然语言处理已广泛应用于金融领域。

当情感分析应用于专业语言领域(例如金融)时, 任务会更具挑战性。国内外学者在金融文本情感分析领域的研究方法分为 3 种: 基于词典的方法^[1]、基于机器学习的方法^[2]和基于深度学习的方法^[3]。基于词典的方法以情感词典作为情感倾向的判断依据, 虽然对情感倾向的判断较为简单, 但对词典的依赖较高, 且目前缺乏公开的金融领域情感词典。基于机器学习的方法利用金融文本特征进行情感分析, 但是分析效果受训练语料影响较大。目前, 基于深度学习的方法在计算机视觉、自然语言处理等领域应用广泛^[6], 然而由于金融领域包含大量专业词汇, 通用的情感分析模型并不适用, 且难以提取到文本更深层次的情感特征。

针对现有方法的不足, 本文提出一个基于 FinBERT-RCNN 的金融文本情感分析模型, 该模型采用 FinBERT 提取金融文本上下文联系和语义信息, 情感分类任务的关键在于特征包含的信息, 得到金融文本的语义特征后, 引入 RCNN 模型提取关键信息进行情感分类。通过对比分析实验, 验证了该模型的有效性。

2 相关工作

2.1 基于词典的方法

情感分析是一种分类任务, 旨在识别文本中表达的是积极情绪还是消极情绪。在情感分析的早期研究中, 大多数研究使用基于词典的方法。情感词典的方法是首先人工构造一个情感词典, 词典的内容足够丰富, 就可以获得较好的情感分析效果。Loughran 等^[9]基于词典的方法进行金融文本分析的工作。但是依靠手工提取特征的情感分析方法, 在面对复杂的文本特征时, 仍存在一定的局限性。并且维护词典需要很大的成本^[11], 通用情感词典通常无法考虑特定领域的内涵和对先验知识的大量引用。

2.2 基于机器学习的方法

基于机器学习的方法,通过选择特征项、构造特征对文本数据进行特征提取,实现利用机器学习模型进行情感分类。Alkubaisi 等^[8]利用一种混合朴素贝叶斯分类器的方法对 Twitter 股票评论数据集进行情感分析,并取得了较高的情感分类准确率。Yazdani 等^[10]讨论了分别利用二进制编码、词频和词频-逆文本频率指数对金融新闻文章进行特征提取,并利用具有不同核函数的支持向量机作为分类器对金融文本进行情绪分类,实验结果表明,特征选择和特征加权在情感分类中起到重要作用。Salles 等^[11]针对股票评论数据具有高维噪声的特点,提出延迟随机森林分类模型,通过最近邻样本投影得到与待分类样本相似的特征,具有较高的分类准确率。张对^[14]通过对新浪股吧中的股评数据进行分词和过滤停用词,将剩下的名词、动词和副词等作为特征项进行词频统计,然后利用 SVM 分类器对股评数据进行情感分类。但是,基于机器学习的情感分析方法对特征项选取和特征构建的质量依赖较高,且缺少优质的股票评论标注数据集用于训练。

2.3 基于深度学习的方法

基于深度学习的方法,首先把输入映射到不同的特征空间来进行特征提取,然后持续地修正神经网络权重以学习文本特征。Jiang 等^[12]将基于门控和关注机制的双向长短期记忆神经网络模型用于处理股票新闻和微博的情感分析任务,首先利用门控机制整合字符级别和单词级别的嵌入,然后利用双向长短期记忆神经网络组件将目标相关信息嵌入语句,最后利用线性回归层进行情感分类。Rao 等^[14]提出将基于两个隐藏层的长短期记忆网络模型用于股票评论情感分析,其中,第一层用于学习句子语义,第二层用于学习句子关系。Sohangir 等^[15]将长短期记忆网络和卷积神经网络这两个模型进行对比,结果表明,卷积神经网络在股票评论数据集上表现更好。Akhtar 等^[6]利用卷积神经网络和词向量模型对股评数据进行情感分析,利用词向量模型将股评数据转化为向量,并通过卷积神经网络得到分析结果。直接采用卷积神经网络对股评数据进行情感分析能够实现自动提取特征,修正学习输出,但缺少优质的股票评论标注数据集用于卷积神经网络的训练,且卷积神经网络无法处理股评数据中的专业词汇。

Liu 等^[10]提出基于 BERT 的预训练模型 FinBERT,并采用一个大型通用的金融语料集对其进行训练。FinBERT 首先在 TRC2-financial 语料库上训练语言模型,然后利用其权重初始化金融评论情感分析模型。FinBERT 由编码器和解码器堆叠形成的 Transformer 组成,每个 Transformer 利用编码器的多头自注意力机制将任意位置的词之间建立联系。FinBERT 能够解决目前金融领域缺乏优质股票评论数据集的问题,且能够解决无法处理金融专业领域词汇的问题,但是使用 FinBERT 仅能提取到股票评论的语义特征,而缺少股票评论的局部特征和空间特征。

针对现有模型存在缺少优质标注数据集用于模型训练、无法处理专业词汇、特征提取过于单一等问题,本文提出一种基于 FinBERT-RCNN 的金融文本情感分析模型,利用 FinBERT 预训练模型提取语义特征并使用循环卷积神经网络模型 RCNN 来捕捉数据的局部特征和空间特征,在对比实验中取得了较好的结果。

3 基于 FinBERT-RCNN 的金融文本情感分析模型

3.1 模型设计

FinBERT 预训练模型花费了大量算力、大量金融数据进行了训练,更适合金融相关任务,因此使用 FinBERT 进行嵌入,可以更好地捕获语义特征并提供给模型;另一方面,循环卷积神经网络(RCNN)能够尽可能多地捕获上下文信息,提取文本中关键特征,因此,可以采用该模型来进一步提取文本特征。FinBERT-RCNN 模型的具体结构如图 1 所示。

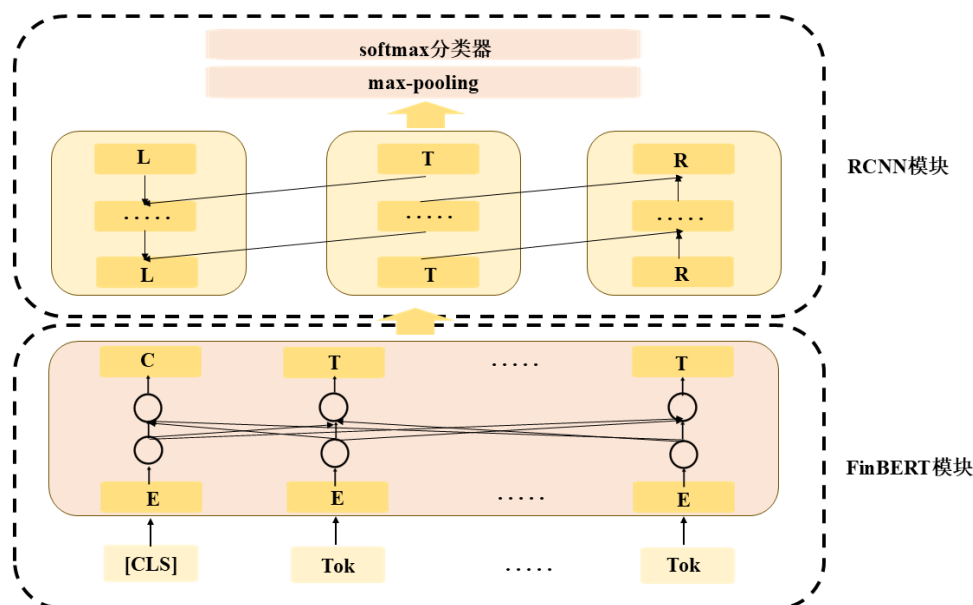


图 1 FinBERT-RCNN 模型结构图

3.2 FinBERT 模块

为了更有效地提取金融文本的语义特征，首先利用 FinBERT 预训练模型对预处理后的金融文本数据进行嵌入处理，即 Token Embeddings、Segment Embeddings 和 Position Embeddings(见图 2)。

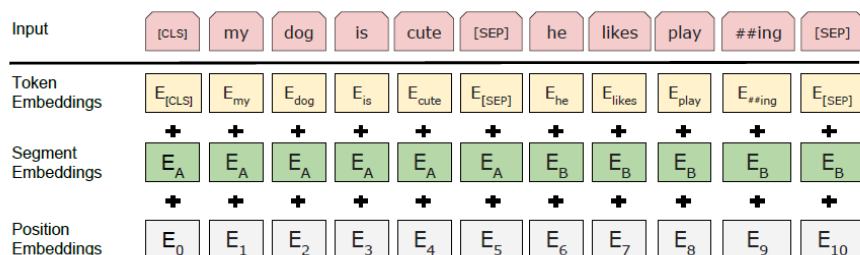


图 2 FinBERT 输入表示

- (1)Token Embeddings: 将词转为固定维度的向量；
- (2)Segment Embeddings: 辅助 FinBERT 区别句子对中的两个句子的向量表示；
- (3)Position Embeddings: 辅助 FinBERT 学习到输入的顺序属性。

最后，将 3 个向量的加和作为模型输入，可以保留更多有效信息及情感倾向，利于后续任务，最终输入包含金融文本数据的词向量、文本向量、位置向量。

$$E_{word} = E_{token} + E_{seg} + E_{pos} \quad (1)$$

将得到的结果作为输入向量，通过一个包含了多层多个 Transformer 的特征提取器可以提取到特征信息 $T_1 T_2 \dots T_N$ 。

3.3 RCNN 模块

情感分类任务的关键在于特征包含的信息。虽然 FinBERT 模型可以更好地处理上下文联系和语义信息，但是其缺少关键信息的提取。直接将 FinBERT 获得的特征用于分类的效果有限，其模型性能仍有提高的空间。利用 FinBERT 模型提取语义特征后，再结合其他神经网络模型可以进一步提升整体分类精度。因此，在模型中引入 RCNN，可以更好地提取特征信息进行情感分类。

RCNN 模型将传统卷积神经网络的卷积替换为内部具有递归结构的循环卷积层，并按前馈连接建立深度网络结构，循环卷积层结构如图 3 所示。

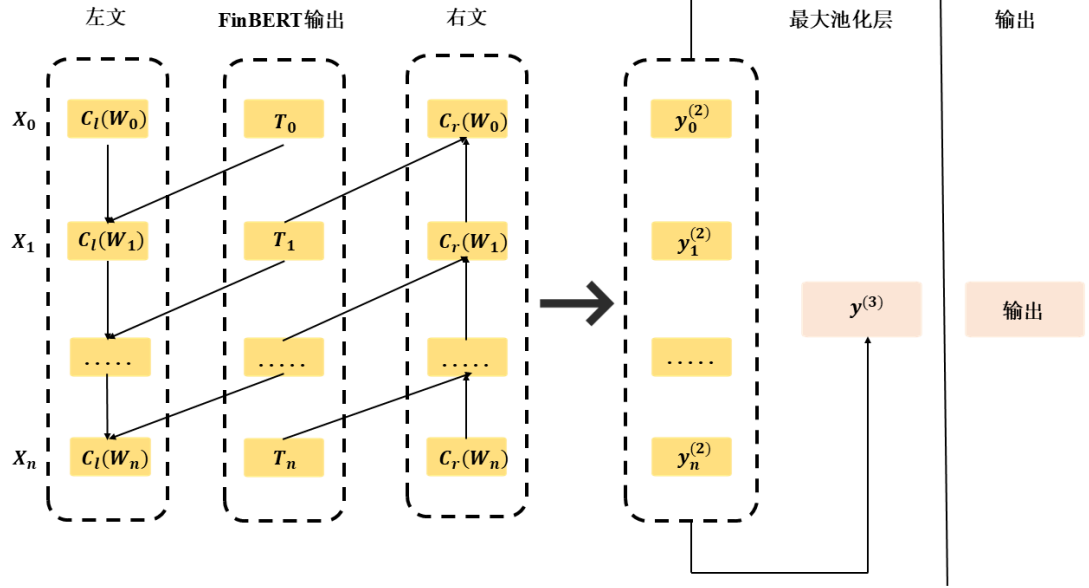


图 3 循环卷积层结构

每个单词会生成两个向量，首先产生从左到右的学习，再产生从右到左的学习。 x 包含了字符及其上下文向量，在 RCNN 中将其输入池化层，通过 max-pooling 取得文本中的最具有代表性的关键特征，再用全连接层输出，最后通过 softmax 函数获得分类的结果。

4 实验与分析

4.1 数据集

选取公开数据集 Financial PhraseBank 对提出的模型和其他现有模型进行实验。数据集中训练集共约 10 000 条、验证集约 1000 条、测试集约 1000 条，包括文本内容、情绪极性(积极、消极、中立)。

4.2 评估指标

实验使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 分数(F1-Score)作为衡量模型分类性能的标准，计算公式如下所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

其中，TP(True Positive)表示被正确预测的正例，即该数据的真实值为正例，预测值也为正例的情况；FP(False Positive)表示被错误预测的正例，即该数据的真实值为反例，但被错误预测成了正例的情况；FN(False Negative)表示被错误预测的反例，即该数据的真实值为正例，但被错误预测成了反例的情况；TN(True Negative)表示被正确预测的反例，即该数据的真实值为反例，预测值也为反例的情况。

4.3 实验环境

实验基于 Pytorch 框架，训练时所使用的 GPU 为 NVIDIA GeForce RTX 3090，操作系统为 Ubuntu 24.02 64bit。训练时使用的模型参数如表 1 所示。

表 1 模型参数

参数	数值
词向量维度	768
Dropout	0.1
Learning rate	5e-5
Batch size	128
Pad size	32

4.4 对比实验

为了验证提出的模型在金融领域情感分析上的效果，本节对以下模型进行了对比实验，选择了 5 种针对金融文本情感倾向分析的方法模型与本文提出模型进行对比，包括基于词典的方法^[9]、SVM^[13]、TextCNN^[15]、BERT^[11]和 FinBERT^[16]。

表 2 不同方法的测试结果

方法	准确率	召回率	F1 值
基于词典的方法	0.4849	0.4515	0.4459
SVM	0.7999	0.7906	0.8007
TextCNN	0.8103	0.8082	0.8071
BERT	0.8554	0.8464	0.8449
FinBERT	0.8935	0.8542	0.8898
FinBERT-RCNN	0.9512	0.9356	0.9509

表 3 FinBERT-RCNN 的详细结果

Label	Precision	Recall	F1-Score	Support
Positive	0.9600	0.9781	0.9689	319
Negative	0.9590	0.8931	0.9249	131
Neutral	0.8923	0.9355	0.9134	62
Accuracy			0.9512	512
Macro Avg	0.9371	0.9356	0.9357	512
Weighted Avg	0.9516	0.9512	0.9509	512

表 4 FinBERT 的详细结果

Label	Precision	Recall	F1-Score	Support
Positive	0.8900	0.9720	0.9292	3537
Negative	0.9272	0.7083	0.8031	1457
Neutral	0.8610	0.8824	0.8716	723
Accuracy			0.8935	5717
Macro Avg	0.8927	0.8542	0.8680	5717
Weighted Avg	0.8958	0.8935	0.8898	5717

由表 2、表 3、表 4 可知，本文提出的 FinBERT-RCNN 模型在准确率、召回率和 F1 值方面均显著优于其他模型。FinBERT-RCNN 模型的准确率达到 0.9512，相比 FinBERT 的

0.8935 提升了约 6.44%，表明 FinBERT-RCNN 在综合情感分类能力上表现更加优秀；基于词典的方法效果最差，4 个指标均不到 0.5，究其原因其没有公开全面的高质量股票情感词典。SVM 的实验结果与基于词典的方法相比准确率提升大约 0.3，但由于其依赖特征质量，且需要已标注的数据，所以仅优于基于词典的方法。TextCNN 相较于基于词典的方法和 SVM 在 4 个评价指标上均有大幅提升，但稍逊于 BERT 和 FinBERT。综合上述分析，RCNN 模型在捕捉文本序列中的上下文信息方面表现卓越，结合 FinBERT 的预训练语义表示能力，能够更好地挖掘金融领域文本中的深层次语义和依赖关系，特征学习能力得到增强，对不平衡数据的适配性更优。

5 结束语

本文提出了一种应用于金融领域的情感分类模型，该模型集成了 FinBERT 和 RCNN 两个模块，通过 RCNN 模块对 FinBERT 嵌入层的输出进行特征提取，捕获了文本序列中的关键特征，提高了模型情感分类的准确性。实验证明，提出的金融领域情感分析模型与其他模型相比，在性能上获得了明显提升。

参考文献

- [1] Hudson R, Muradoglu YG. Personal routes into behavioural finance [I. Review of Behavioral Finance, 2020, 12(1): 1-9.
- [2] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1-135
- [3] Maqsood H, Mehmood I, Maqsood M, et al. A local and global event sentiment based efficient stock exchange forecasting using deep learning [J]. International Journal of Information Management, 2020, 50: 432-451.
- [4] Jiang M, Wang J, Man L, et al. An effective gated and attention-based neural network model for fine-grained financial target-dependent sentiment analysis [C]// International Conference on Knowledge Science, Engineering and Management, 2017: 42-54.
- [5] Rao G, Huang W, Feng Z, et al. LSTM with sentence representations for document-level sentiment classification [I]. Neurocomputing, 2018, 308: 49-57.
- [6] Sohagir S, Wang D, Pomeranets A, et al. Big[12] Data: deep learning for financial sentiment analysis [J]. Journal of Big Data, 2018, 5(1): 1-25.
- [7] Akhtar MS, Kumar A, Ghosal D, et al. A Multilayer Perceptron based Ensemble technique for fine-grained financial sentiment analysis [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 540-546.
- [8] Liu Z, Huang D, Huang K, et al. FinBERT: a pre-trained financial language representation model for financial text mining [C]// Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence, 2020: 4315-4519.
- [9] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey [J]. Ain Shams Engineering Journal, 2014, 5(4): 1093-1113.
- [10] Kim Y. Convolutional neural networks for sentence classification [J]. arXiv Preprint, arXiv: 1408.5882, 2014.
- [11] Liu C, Wang W, Wang M, et al. An efficient instance selection algorithm to reconstruct training set for support vector machine [J]. Knowledge-Based Systems, 2017, 116(1): 58-73.
- [12] Bishop CM. Pattern Recognition and Machine Learning (Information Science and

Statistics)[MSpringer-Verlag New York, Incorporated, 2006.

- [13] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [J]. arXiv Preprint, arXiv:1810.04805.2018.
- [14] Li Y, Ran J. Investor sentiment and stock price premium validation with Siamese twins from China[J]. Journal of Multinational Financial Management.2020:57-58.
- [15] Chen RD, Bao WW, Jin CL. Investor sentiment and predictability for volatility on energy futures markets: evidence from China [J]. International Review of Economics and Finance, 2021,75(2):112-129.
- [16] Yang CP, Wu HH. Investor sentiment with information shock in the stock market [1]. Emerging Markets Finance and Trade, 2021,57(2):510-524.