Emotion Analysis of Industry Research Report Based on the Hybrid Method of BERT and BiLSTM

Qi Luo[1,2], Mu Zhang[1,*]

[1] *School of Big Data Application and Economics, Guizhou University of Finance and Economics, Guiyang (550025), Guizhou, China*

[2] *Guizhou Institution for Technology Innovation & Entrepreneurship Investment, Guizhou University of Finance and Economics, Guiyang (550025), Guizhou, China*

[*] *Corresponding author. Email: zhangmu01@163.com*

ABSTRACT

In this era of information explosion, industry research papers contain a large number of information about the current business situation and profit and loss of enterprises, which reflects the credit risk of enterprises from the side, further affecting the objective evaluation of financial institutions such as banks. In order to improve the accuracy of the emotional analysis of industry research papers, this paper adopts the emotional analysis method based on the combination of BERT and BiLSTM, and selects 100 industry research papers from 38 industries of Dongfang Fortune Network in 2021 as samples for emotional analysis. To better demonstrate the performance of this model, this paper uses the SnowNLP model to compare with it. The results show that the accuracy, recall and F1 values based on the mixed model of BERT and BiLSTM are 85.42%, 82% and 83.68% respectively. The accuracy, recall and F1 value of SnowNLP model are 75%, 78% and 76.47% respectively. It can be concluded that the performance of the method proposed in this paper is good, and it has certain validity for the sentiment analysis of industry research papers, and has certain value in helping to predict the future development trend of the industry.

Keywords: *BERT, BiLSTM, Industry Research Report, Affective Analysis.*

基于 BERT 与 BiLSTM 混合方法的行业研报情感分析

罗 琦[1,2]，张 目[1,*]

[1] *贵州财经大学 大数据应用与经济学院，中国 贵州 贵阳 550025*

[2] *贵州财经大学 贵州科技创新创业投资研究院，中国 贵州 贵阳 550025*

[*] *通讯作者. 电子邮箱：zhangmu01@163.com*

摘要

在这个信息大爆炸的时代，行业研报中包含着大量企业经营现状以及盈亏的信息，这些信息从侧面反映出企业的信用风险，进一步影响着银行等金融机构展开客观评价。为了提高行业研报情感分析的准确度，本文采用基于 BERT 和 BiLSTM 相结合的情感分析方法，选取 2021 年东方财富网 38 个行业共 100 篇行业研报

S. Qiao et al. (Eds.): RAC 2022, AEBMR 248, pp. 170–176, 2023.
https://doi.org/10.2991/978-94-6463-194-4_24

作为样本，对其进行情感分析。为更好展示该模型的性能，本文采用 SnowNLP 模型与之进行对比。其结果显示，基于 BERT 与 BiLSTM 混合模型的精确率、召回率和 F1 值分别为 85.42%、82%和 83.68%；SnowNLP 模型的的精确率、召回率和 F1 值分别为 75%、78%和 76.47%。可以得出，本文所提出的方法性能较好，对行业研报文本情感分析具有一定的有效性，在帮助对行业未来的发展趋势进行预测上具有一定的价值。

关键字：*BERT, BiLSTM, 行业研报, 情感分析.*

1. 引言

本文通过对行业研报文本中提取的研究人员的判断和观点语句，形成情感分析数据集，通过 BERT 和 BiLSTM 混合的情感分析方法，判断训练集中语句的情感极性。本文的创新性在于：第一，创新性地将 BERT 模型应用行业研报情感分析中，能更有效地进行文本特征提取；第二，创新性地将 BERT 与 BiLSTM 相结合进行行业研报情感分类，其结果表明该模型有效且实用，可以帮助对行业未来的发展趋势进行预测，为企业制定行业市场战略、预估行业风险提供参考。

2. 国内外研究现状

随着人工智能的发展，使用自然语言处理技术的文本数据的情感分析已广泛运用于多个领域，例如水利施工[1]、网络舆情检察[2]、疾病预测[3]以及金融领域[4]等。

早期的情感分析研究主要使用基于情感词典的方法。王文韬和张士豹[5]通过新浪微博中有关新冠疫情话题的评论数据，结合情感词典和支持向量机的方法构建情感分类模型，最终预测出微博网民在新冠疫情期间的感情以积极为主。Chayan Paul 和 Pronami Bora[6]通过网络上大量印度超级联赛的评论信息，结合词典的方法，来研究用户对联赛的情绪。但是基于词典的情感分析方法对情感词典的依赖性极高，不同词典对同一文本的情感分析效果不同，其精确度受限于情感词典和情感判断规则的搭配，随着数据量的不断增加变化，基于词典的方法不能较好完成文本的情感分析问题。

现有研究证明，基于机器学习的方法在文本情感分析的正确率上高于基于词典的方法[7]。尚永敏和赵榆琴[8]使用朴素贝叶斯、支持向量机和 SnowNLP 方法对文本数据进行情感分析，通过对比三种机器学习方法，最终实现基于 SnowNLP 和 LDA 的在线情感分析方案。Kamal 等人[9]通过提出结合基于规则和机器学习方法的两种方法的情绪分析系统来辨别特征与意见作用于其情感极性，成功利用用户对不同种类电子产品的评价实现了划分用户的情感极性。但是这些基于机器学习的方法进行的情感分析由于需要手动提取文本的数据的情感特征，当文本的特征比较复杂或者文本的数据量比较庞大时，仍然具有一定的局限性。

目前在处理任务上获得最领先成果的情感分析方法是在论文《Pre-training of Deep Bidirectional Transformers for Language Understanding》中提出的 BERT 模型[10]，该模型对于 11 种自然语言的任务处理方面均处于领先地位。BERT 模型由于其双向 Transformer[11]结构，该结构的自注意力机制可以更好的进行特征提取以及解决长文本所带来的上下文之间的语义依赖问题。黄建民等[12]研究在 BERT 模型的基础上加入两种模块：分别为并行聚合模块和层次聚合模块,这两个模块主要用于方面抽取和方面情感分类；Bedi Jatin 与 Toshniwal Durga[13]的研究阐述了一种依据 BERT 的情感分类和投诉分类模型,提高了两种分类的精度。

3. BERT 训练模型的介绍

BERT（Bidirectional Encoder Representation from Transformers），是一种为不同的自然语言（Natural Language Processing, NLP）任务提供支持的通用的新型语言表征模型。BERT 是 2018 年谷歌发布的模型，该模型中的 Transformer 层采用双向编码器表示，其先进性在于使用 Masked Language Model（MLM）和 Next Sentence Prediction（NSP）的新预训练任务[14]。

BERT 采用双向 Transformer 提取句子特征，其

结构更强大，学习能力更强。Transformer 本质上是一个编码器—解码器模型，其核心为自注意力机制（Self-Attention）[11]，其作用是在大量信息中挑选出当前任务更需要的关键信息。

4. 双向长短期记忆模型（BiLSTM）的介绍

双向长短期记忆网络模型（BiLSTM）主要由两个方向相反的 LSTM 网络组成，即一个向前的 LSTM 网络按照从前往后的顺序来进行文本信息的读取，另一个向后的 LSTM 则按照从后往前的顺序来进行文本信息的读取，最后将两个方向所得到的输出信息进行连接，便可以得到同时具备前后两个方向信息的文本特征。其延续了 LSTM 的优点，即没有梯度消失和梯度爆炸的问题，同时也解决了 LSTM 只可以向前学习不能往后学习的问题，其有效的联系文本前后的含义，更好的获得双向的句子依赖。

5. 基于行业研报文本情感分析的模型设计

5.1 行业研报文本数据预处理

本文研究需要获取各类行业研报，并对其进行预处理。本文所采用的文本数据为东方财富网所发布的 38 个行业的研报，其中 2021 年之前发布的研报共选取 800 篇作为训练集；2021 年到 2022 年发布的行业研报，每个行业选取两到三篇，共 100 篇作为测试集。

行业研报文本数据的预处理主要是对原始数据进行清洗、去除噪声及无关的内容，得到高质量的数据，使之后的情感分析结果更为准确。包含以下步骤：从东方财富网上获取行业研报文本数据；取出研报中的非法字符和无用词语等，并去除文本中的空格；提取文本分类结果，将积极态度研报打上标签 1,将消极态度研报打上标签 0。

5.2 基于 BERT-BiLSTM 的行业研报文本情感分析模型

BERT 预训练语言模型能构建精确的文本向量表示，学习特征权重分布，加强对有效信息的关注；且 BiLSTM 模型在处理时序数据、提取上下文文本特征上具有突出优势[16]。因此，本文结合 BERT 和 BiLSTM 模型构建行业研报文本情感分析模型。其结构如图 1 所示，其主要包含以下四层：输入层，主要进行文本数据的预处理，通过查询字向量表将原始数据中的每个字变成一维字向量输入到模型中；基于 BERT 的词嵌入层，利用 BERT 中的 Transformer 将词向量、句子向量和位置向量相加，得到输入文本对应的融合全文语义信息后的向量表示；BiLSTM 特征抽取层，利用 BiLSTM 模型进行特征提取和语义编码处理，获得整篇行业研报的文本特征；情感计算层，将 BiLSTM 特征抽取层获得的句子表示经全连接层，使用 Softmax 激活函数对整篇行业研报的情感进行计算，最终获得这篇行业研报的情感分类。

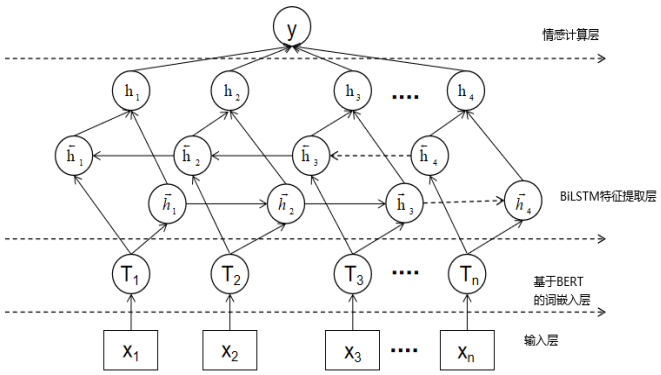

图 1 基于 BERT 和 BiLSTM 的模型结构

（1）BERT 获得向量表示。本文选择更适合中文任务的 BERT-base-Chinese 模型作为预训练模型，行业研报文本经过分词后输入 Encoder 编码模块得到对应的词序转化后的索引，而后将每条行业研报文本的索引输入 BERT 获得每个字的词向量。本实验中，将行业研报文本的最大长度限度为 200 个字，对长度超过 200 字的文本进行截断，少于 200 字的文本用 0 填充，同时在输入文本的开头和结尾部分分别添加[CLS]和[SEP]标识符。对于 BERT 模型，其输入由三个部分相加得到：分别是字向量、段向量和位置向量。

在获得上述行业研报文本的词向量后，将其输入到 BERT 模型中，BERT 模型同时进行 Masked LM 和 NSP 两个与训练任务后，可以从大量的文本信息中学习到字符级、词语级和语句间关系的特征。

预训练任务 1：Masked LM

以输入文本“工业机器人产量创单月新高，行业高景气度延续”为例，BERT 在进行 MLM 预训练任务过程中，会随机选择 15%的字词用于预测。具体来说，就是输入的文本由 80%的机率变成“工业机器人产量创单月新[MASK]，行业[MASK]景气度延续”，[MASK]字符表示“高”被遮盖，需要利用 BERT 模型对遮盖部分进行预测；有 10%的概率输入的文本会变成“工业机器人产量创单月新低，行业低景气度延续”，即把“高”替换成了其他字词，比如“低”；有 10%的概率输入的文本保持不变。

预训练任务 2：Next Sentence Prediction

在进行 NSP 预训练任务时，BERT 会挑选一半的训练数据为连续的语句对，另一半则为不连续的语句对，然后利用 BERT 对这些文本数据进行监督训练，从而学习到句子预句子之间的关系。

（2）BiLSTM 提取特征。本文在行业研报文本的语义信息提取时采用 BiLSTM 网络，将前向传播的向量与反向传播的向量进行连接，以此来同时获得上下文语义信息。具体来说，将 BERT 的输入中字符[CLS]对应的输出 C 乘以权重 W，作为 BiLSTM 网络的输入，其计算公式（1）如下：

$$a_i = g(W_a C + b) \tag{1}$$

然后，模型把输入向量输入隐层中，BiLSTM 在两个不同方向的隐层上进行计算，最后把两个方向的结果拼接输出，即 $h_i = \vec{h}_i + \overleftarrow{h}_i$。$\vec{h}$ 表示前向传播隐藏层向量，$\overleftarrow{h}$ 表示后向传播隐藏层向量。隐藏层的激活函数采用 tanh，其计算过程如下：

$$h_i^d = \sigma(W_h^d a_i + U h_{i-1}^d + b_h^d) \tag{2}$$

其中 W_h^d 表示第 d 和索引对应的 a_i 权重矩阵，U 是 $i-1$ 时刻隐藏层输出 h_i^d 对应的权重矩阵，d 表示隐藏层的连个不同方向，b_h^d 代表第 d 个索引对应的偏置向量。最后，将一层的所有向量 h_i^d 进行拼接，作为整个句子的特征向量表示。

（3）情感计算。为了对行业研报文本进行情感分类，本文把 BiLSTM 输出的特征向量经过一个全连接层后，输入 Softmax 激活函数对其进行情感分类结果的预测。对于每篇行业研报文本，模型最后都会输出一个向量，用来表示该篇文本属于正面或负面的概率

$$p(y \mid H, W_c, b_c) = soft\max(W_c H + b_c) \tag{3}$$

其中，p 为情感分析为正面或负面的概率，H、W_c 和 b_c 为 BiLSTM 网络输出层参数。

6. 实证分析

6.1 数据来源

本文的数据集分为训练集和测试集两部分，来源于东方财富网行业研报文本数据。本文选取东方财富网中的 38 个行业 2021 年以前所发布的共 800 篇行业研报文本数据组成训练集；然后选取东方财富网中的 38 个行业 2021 年到 2022 年的行业研报各两到三篇，共 100 篇行业研报文本数据组成测试集,其中正向情绪为 50 条，负向情绪为 50 条。

6.2 基准模型

为更好地评价本文所提出的基于 BERT 与 BiLSTM 结合的情感分析模型，本文选择 SnowNLP 模型进行对比，以此来验证本文所提出的模型情感分析的结果更精确。

6.3 评价指标

为了有效地评估基于 BERT 与 BiLSTM 结合的情感分析模型的性能，本文选用分类任务中常用的精确率（Precision）、召回率（Recall）和 F1 值（F1 Score）作为情感分析效果的评价指标。其计算公式如式（4）

到式（6）所示：

$$\Pr ecision = \frac{TP}{TP + FP} \tag{4}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2\Pr ecision \times \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \tag{6}$$

其中，TP 表示预测为正向且正确的样本数量，TN 表示预测为负向且正确的样本数量，FP 表示预测为正向但为负的样本数量，FN 表示预测为负向但为正的样本数量。

6.4 实验结果

为了验证文本提出的基于BERT与BiLSTM结合的情感分析模型的可行性和有效性，本文选取 2021 年到2022年之间东方财富网38个行业的行业研报文本信息所组成的 100 条测试集进行情感分类预测，并利用 SnowNLP 模型进行对比实验。其结果如表 1、表 2 所示。

从表 2 可以看出，总体来讲，BERT-BiLSTM 模型的情感分析结果明显好于 SnowNLP 的情感分析结果。BERT+BiLSTM 模型的精确率为 85.42%，而 SnowNLP 模型的精确率为 75%，BERT+BiLSTM 模型的召回率为 82%，而 SnowNLP 模型的召回率为 78%，BERT+BiLSTM 模型的 F1 值为 83.68%，而 SnowNLP 模型的 F1 值为 76.47%。BERT+BiLSTM 模型各方面的数据都优于 SnowNLP 模型。可以得出，本文提出的基于BERT与BiLSTM结合的情感分析模型性能较好，适用于各领域情感分析研究。

7. 结论

本文针对行业研报情感分析任务，为了获得更多的行业研报文本特征，提出了基于 BERT 和 BiLSTM 混合模型。为验证提出模型的先进性和有效性，用 SnowNLP 模型进行对比试验，试验结果表明，本文提出的模型在行业研报情感分析任务上取得更高的精确率、召回率和 F1 值。本文提出模型可广泛运用各领域文本情感分析任务上。

表 1 两个模型情感分类预测结果

	BERT 和 BiLSTM 结合的情感分析模型		SnowNLP 模型	
	预测值		预测值	
真实值	1	0	1	0
1	41	9	39	11
0	7	43	13	37

表 2 两个模型性能结果

模型	**精确率**	**召回率**	**F1 值**
BERT+BiLSTM	85.42%	82%	83.68%
SnowNLP	75%	78%	76.47%

致谢

本研究得到国家自然科学基金地区项目“基于文本信息的科技型中小企业信用风险识别机理研究”（71861003）的资助。

参考文献

[1] 刘婷,张社荣,王超,李志竑,关炜,王枭华.基于BERT-BiLSTM混合模型的水利施工事故文本智能分析[J/OL].水力发电学报:1-13[2022-06-25]. http://kns.cnki.net/kcms/detail/11.2241.TV.20220303.1734.002

[2] 刘继,顾凤云.基于BERT与BiLSTM混合方法的网络舆情非平衡文本情感分析[J].情报杂志,2022,41(04):104-110.

[3] 龚汝鑫,余肖生.基于BERT-BILSTM的医疗文本关系提取方法[J].计算机技术与发展,2022,32(04):186-192.

[4] 朱鹤,陆小锋,薛雷.基于BERT的金融文本情感分析模型[J/OL].上海大学学报(自然科学版):1-15[2022-06-25].http://kns.cnki.net/kcms/detail/31.1718.n.20210616.1757.002.html

[5] 王文韬,张士豹.基于情感词典和SVM的微博网民情感分析[J].现代信息科技,2021,5(24):24-27+31.DOI:10.19850/j.cnki.2096-4706.2021.24.007.

[6] A Dictionary Based Analysis of User’s Sentiment Regarding Indian Premier League[J]. International Journal of Innovative Technology and Exploring Engineering,2019,8(11): 963-965.

[7] LI, F. The Information Content of Forward‐Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach [J]. Journal of Accounting Research, 2010, 48: 1049-1102.

[8] 尚永敏,赵榆琴.基于机器学习的在线评论情感分析与实现[J].大理大学学报,2021,6(12):80-86.

[9] KAMAL A,ABULAISH M.Statistical features identification for sentiment analysis using machine learning tech niques[C]//International Symposium on Computational& Business Intelligence, 2013: 178-181.

[10] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding [J], ArXiv181004805 Cs, 2018,19: 4171-4186.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need" [J], CoRR, 2017: 1-15.

[12] 黄建民,李强,王雪绒,李聪聪.基于BERT融合多模块的方面级情感分析[J].井冈山大学学报(自然科学版),2021,42(06):64-68.

[13] Bedi Jatin and Toshniwal Durga. CitEnergy : A BERT based model to analyse Citizens ’ Energy-Tweets[J]. Sustainable Cities and Society, 2022(80).

[14] 张枫叶. 基于BERT和LDA的阅读软件评论情感分析研究[D].曲阜师范大学,2021.DOI:10.27267/d.cnki.gqfsu.2021.000073

[15] 许雪晨,田侃.一种基于金融文本情感分析的股票指数预测新方法[J].数量经济技术经济研究,2021,38(12):124-145.DOI:10.13653/j.cnki.jqte.2021.12.009.