Advances in Applied Mathematics 应用数学进展, 2023, 12(3), 1090-1099
Published Online March 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.123111

Hans汉斯

基于机器学习与先进transformer模型的情感预测

孙　睿[1]，周艳聪[2*]

[1]天津商业大学理学院，天津
[2]天津商业大学信息工程学院，天津

收稿日期：2023年2月17日；录用日期：2023年3月13日；发布日期：2023年3月21日

摘　要

本文立足于针对文本的情感分析，以Yelp数据集为例进行评估。Yelp评论的评级预测可以通过多种方式进行，如情绪分析和五星评级分类。在本文中，我们将基于评论文本对餐馆的评级进行预测。在分析了原始数据分布之后，首先创建了一个平衡的训练子数据集，后分割数据集、提取特征，同时应用朴素贝叶斯和Logistic回归两种机器学习方法和基于transformer的BERT、DistilBERT和RoBERTa三种深度学习模型进行评估比较。从训练时间和训练效果两个方面给出结果，为读者提供实际的选择依据。

关键词

朴素贝叶斯，Logistic回归，BERT，DistilBERT，RoBERTa

Emotion Prediction Based on Machine Learning and Advanced Transformer Model

Rui Sun[1], Yancong Zhou[2*]

[1]Department of Science, Tianjin University of Commerce, Tianjin
[2]Department of Information Engineering, Tianjin University of Commerce, Tianjin

Received: Feb. 17th, 2023; accepted: Mar. 13th, 2023; published: Mar. 21st, 2023

Abstract

Based on the emotional analysis of the text, this paper takes Yelp data set as an example to eva-

*通讯作者。

文章引用: 孙睿, 周艳聪. 基于机器学习与先进 transformer 模型的情感预测[J]. 应用数学进展, 2023, 12(3): 1090-1099.
DOI: 10.12677/aam.2023.123111

luate it. The rating prediction of Yelp reviews can be made in many ways, such as sentiment analysis and five-star rating classification. In this paper, we will predict the rating of restaurants based on the review text. After analyzing the distribution of the original data, a balanced training sub-data set is first created, then the data set is segmented and features are extracted. At the same time, two machine learning methods, naive Bayes and Logistic regression, and three deep learning models based on transformer, BERT, DistilBERT and RoBERTa, are applied to evaluate and compare. The results are given from two aspects: training time and training effect, which provides practical basis for readers to choose.

Keywords

Naive Bayes, Logistic Regression, BERT, DistilBERT, RoBERTa

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

情感分析是分析一段文本并确定表达者态度的过程，人们可以根据文本中所呈现的内容来预测情感倾向，从而做出判断。随着对文本分析的技术方法不断发展，对情感预测的准确率也在不断提高，放在商业环境中，情感预测被广泛用于评估客户满意度、进行市场调查和监测品牌声誉等领域[1]。如今数字媒体盛行，对企业在线评论的重要性越发不容忽视，客户可以发布他们对企业的评论，其他潜在客户或店主也可以查看这些评论。来自顾客的正面反馈可能会提高业绩，而负面的反馈可能会产生相反的后果。本文旨在一个公开的大型数据集上对现有情感预测方法进行分析比较，验证效果并为企业决策提供依据，为研究者从训练时间和训练效果两个维度上提供参考。

2. 研究现状

作为自然语言处理领域的子研究领域之一，针对文本展开的情感分析进行了大量研究。从方法的演进上大致经过以下几个阶段：构建情感词典、基于机器学习的方法和基于深度学习的方法[2]。随着研究的深入，几种深度学习方法进入研究视野，主要分为四类：基于递归神经网络、基于卷积神经网络、基于记忆网络、基于自我注意机制[3]。近年来，基于 transformer 模型的方法更是进入了研究视野。Yang 等人[4]提出了对评论文本的深度表示，并增加优化了输入文本的部分，以提高整体性能[5]。提出了情感分类任务的端到端新架构，其灵感来自于最近的多任务联合学习的研究[6]。对分类任务采用了多种特征生成方法和朴素贝叶斯、逻辑回归、支持向量机(SVM)和高斯判别分析等几种机器学习模型进行评估。在测试集上，五星分类的最佳准确率为 64% [7]。一些深度学习模型，如神经网络、递归神经网络(RNN)、长短期记忆(LSTM)和双向编码器表示(BERT)也被应用于[8]。随着在该领域的不断研究，目前主要有以下几个挑战：1) 长篇文字和短篇文字中的歧义、组合性、长期依赖性和否定识别仍然是自然语言处理和情感分析中的挑战。2) 跨域任务的性能仍然滞后于在同一领域上训练和测试的效果，需进一步的工作来生成跨不同领域有更好泛化能力的模型。3) 预训练过的语言模型虽已突破了许多 NLP 任务的边界，但对于机构、中小企业和独立研究者无法获得大量资源来说是一个障碍，侧面阻止了前沿深度学习研究的普遍化。

3. 数据与预处理

3.1. 数据集概况

Yelp 是成立于 2004 年的最大的公司之一，旨在发布关于企业的评论，它提供了一个开放的数据集，即 Yelp 开放数据集[9]，它包含了大量关于企业的评价数据，是一个适合个人、教育和学术的标准数据集。在 Yelp 开放数据集的多个任务中，基于他们的评论来预测餐馆的评级是一个基本且重要的任务，这个任务可以被看作是一个多分类问题，即输入文本数据(评论)，输出预测的类(1~5 颗星)。该数据集包括来自 10 个大都市地区的 209,393 家企业的 8,021,122 条评论，这些数据以 JSON 文件为结构进行存储，包括商家、评论、用户和照片等字段。本文将只用到商家数据和评论数据。

3.2. 数据准备

3.2.1 创建子数据集

为了获得更准确的评级预测，本文将只使用餐厅类别中的企业数据。从 business.json 中获取了 52,268 个餐厅的商家 ID，从 review.json 中获取了 4,724,471 条评论。部分数据样本如表 1 所示。同时，每个评级类别的评论数量并不相等，这些数据的分布如图 1 和表 2 所示。可以发现，数据偏向于两个极性，有 67.94%的评论是 4 星和 5 星。这种不平衡的分布在[10]中得到了验证。

Table 1. Example of Yelp Dataset
表 1. Yelp 数据集示例

text	stars
If you decide to eat here, just be aware it is...	3
Family diner. Had the buffet. Eclectic assortm...	3
Wow! Yummy, different, delicious. Our favo...	5
Cute interior and owner (?) gave us tour of up...	4
I am a long term frequent customer of this est...	1

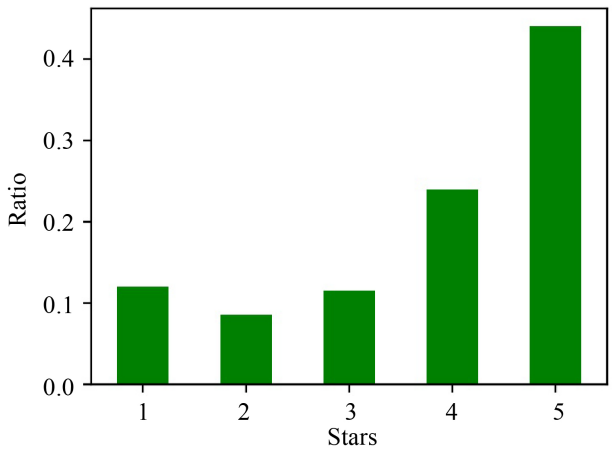

Figure 1. Rating Distribution of Yelp Dataset
图 1. Yelp 数据集评级分布

Table 2. Distribution of Yelp Dataset
表 2. Yelp 数据集分布

stars	评论数	占比
5	2079441	44.01%
4	1130251	23.92%
3	543108	11.50%
2	404486	8.56%
1	567185	12.01%

现为了处理不平衡的原始数据，需重建一个平衡的训练数据集，并以 70:15:15 的比例划分训练集、验证集和测试集。因此，从一个训练数据集的原始数据中重新采样了 100 万条评论，其中每个类别有 20 万条评论(1~5 颗星)。此外，验证集和测试集均有 20 万条评论，但它们大致遵循原始数据的不平衡分布。

3.2.2. 向量化

为了使数据能正确输入模型，首先将文本文档转换为数值数据(矩阵)，以此来构建特征。在[1]中给出了一些详细的文本预处理管道，本文使用 scikit-learn [11]，具体操作如下：

1) 用 Count Vectorizer 和整数型 Tf-Idf Vectorizer 进行单词表示；
2) 使用 unigram 和 bigram；
3) 去停止词；
4) 最低文档频率为 5；
5) 所有单词转换为小写字母；

Count Vectorizer 可以将文本转换为 token 计数矩阵，而 Tf-Idf Vectorizer 将使用 Tf-Idf 来代替 token 计数。Tf-idf 是术语频率 $\mathrm{Tf}(t,d)$ 和逆文档频率 $\mathrm{idf}(t)$ 的乘积：

$$\text{Tf-idf}(t,d)=\mathrm{Tf}(t,d)\times\mathrm{idf}(t) \tag{1}$$

其中 $\mathrm{Tf}(t,d)$ 为文档 d 中术语 t 出现的频率，而 $\mathrm{idf}(t)$ 定义为：

$$\mathrm{idf}(t)=\log\frac{n}{1+\mathrm{df}(t)} \tag{2}$$

其中 n 为本文数据集中的文档总数，$\mathrm{df}(t)$ 为文档集中包含术语 t 的文档数量。对于二进制版本的向量器，所有非零计数都设置为 1。稀疏矩阵也从这些向量器中输出。

4. 实验与结果

4.1. 评估指标

本文将使用以下四种评估指标：

1) 准确率。表示预测正向的样本占总样本个数的比例：

$$\text{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \tag{3}$$

2) 精确率。表示预测为正向的样本在实际也为正向的样本占预测为正向样本的比例：

$$\text{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

3) 加权 F1 值。

$$\frac{1}{\sum_{q\in Q}\left|\hat{y}_q\right|}\sum_{q\in Q}\left|\hat{y}_q\right|F_1\left(y_q,\hat{y}_q\right) \tag{5}$$

其中 Q 是标签的集合， y_q 是带有 q 标签的子集， $\hat{y}_q$ 是预测为 q 的集合，而

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{6}$$

4) 混淆矩阵。以矩阵的形式呈现评估效果，记作 C，行代表数据的真实归类标签，列代表数据被预测的标签， C_{ij} 为实际标签为 i 但被预测为 j 的数量占所有实际标签为 i 的比例。

4.2. 基于经典机器学习方法

文献[11]说明更简单的模型，如逻辑回归和支持向量机，比更复杂的模型 LSTM 和 BERT，能更有效地预测情绪。为验证这一点，本文将经典机器学习方法作为比较基准，选取朴素贝叶斯和 Logistic 回归[12]这两种方法进行实验。其中朴素贝叶斯模型实现最为方便，运行时间最短，其余模型训练时间记录在表 3。在测试集上的准确性、精确性和加权 f1 值如表 4 所示，混淆矩阵分别如图 2、图 3 所示。可以看到 Logistic 回归以 64.23%的准确率优于朴素贝叶斯。

Table 3. Training time record of each model
表 3. 各模型训练时间记录表

模型	运行时间(时：分：秒)
朴素贝叶斯	0:00:03
Logistic 回归	0:08:42
BERT(base, cased)	2:10:37
BERT(base, large)	4:06:11
DistilBERT(base, cased)	0:59:52
RoBERTa(base)	4:47:00

Table 4. Comparison of model effects
表 4. 模型效果比较

模型	Accuracy	Precision	加权 F1 值
朴素贝叶斯	62.09%	64.47%	0.6295
Logistic 回归	64.23%	65.79%	0.6485
BERT(base, cased)	68.11%	69.28%	0.6943
BERT(base, large)	69.13%	69.87%	0.6958
DistilBERT(base, cased)	67.77%	69.43%	0.6901
RoBERTa(base)	69.75%	70.94%	0.6905

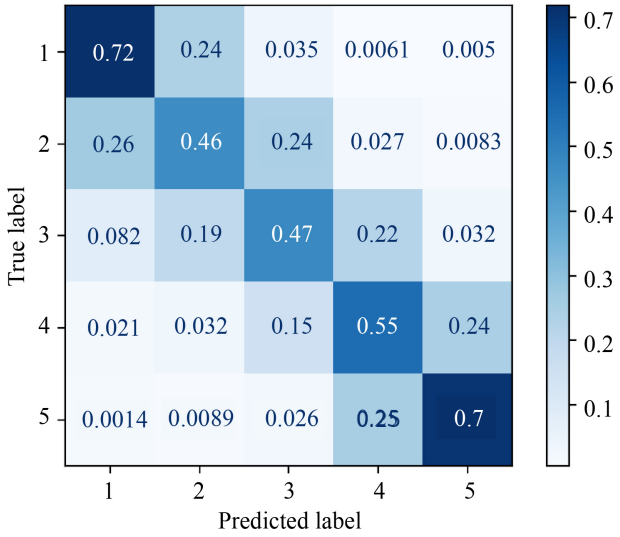

Figure 2. Confusion Matrix of Naive Bayes on Test Set
图 2. 测试集上朴素贝叶斯的混淆矩阵

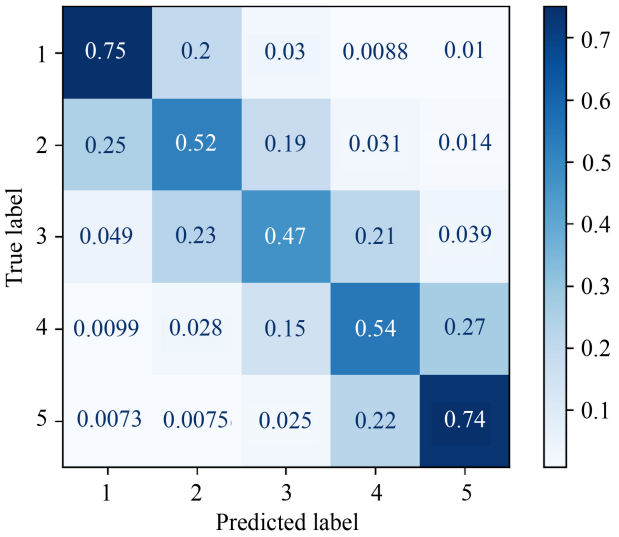

Figure 3. Confusion Matrix of Logistic regression on Test Set
图 3. 测试集上 Logistic 回归的混淆矩阵

4.3. 基于先进 transformer 模型

近年来，基于 transformer 的模型[13]成为重要的情感分析技术，它们在巨大的任务上优于许多其他最先进的方法。在本文实验的第二大部分，我们在 Yelp 评级预测任务中使用了三个基于 transformer 的模型，包括：BERT [14]、DistilBERT [15]和 RoBERTa [16]。需要注意，在该部分由于训练资源的限制，我们将缩小数据集为 10 万条评论，其中每个类别有 2 万条评论(1~5 颗星)，验证集和测试集也均为 2 万条评论。

在进行这部分实验之前，需要对评论的 token 数进行统计分析，如图 4 所示，有 70.47%的评论不超过 128 个标记，92.17%的评论不超过 256 个标记。

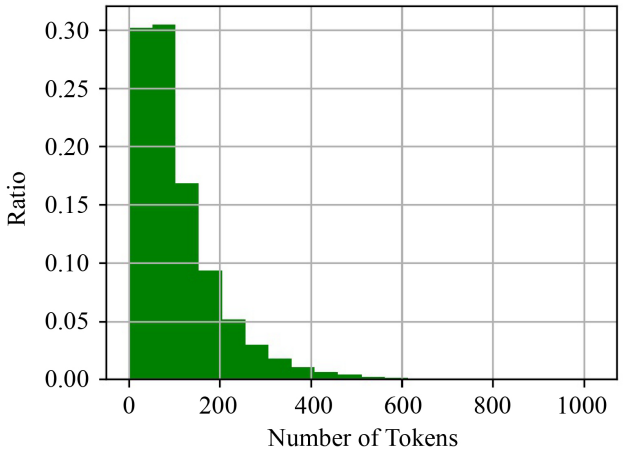

Figure 4. Token number distribution of comment text
图 4. 评论文本的 token 数分布

4.3.1. Deep Bidirectional Transformers (BERT)

BERT 是一种语言表示模型，对未标记文本的左右上下文进行预先训练，用于语言模型和下一个句子预测任务。这样一个模型，当在自然语言文本的大型语料库上进行预训练时，仅用一个额外的输出层进行微调后，可以用于广泛的训练任务。目前有两种 BERT 架构可用：一个较小的 110M 参数的基础 BERT 模型(12 个 transformer 块，768 个隐藏单元，12 个自我注意头)，另一个较大的 340M 参数的 BERT 模型(24 个 transformer 块，1024 个隐藏单元，16 个自我注意头)。本文将对这两种 BERT 模型均进行实验，训练结果如表 4 所示，相应的混淆矩阵如图 5 和图 6 所示。在设置方面，最大序列长度为 128，batch size 为 16。

可以发现，BERT (base, large)要比 BERT (base, cased)准确率高出 1.02%，加权 F1 值提高了 0.0015，这也是符合预期的，前者比后者具有更多的参数，但运行时间多出了一倍。

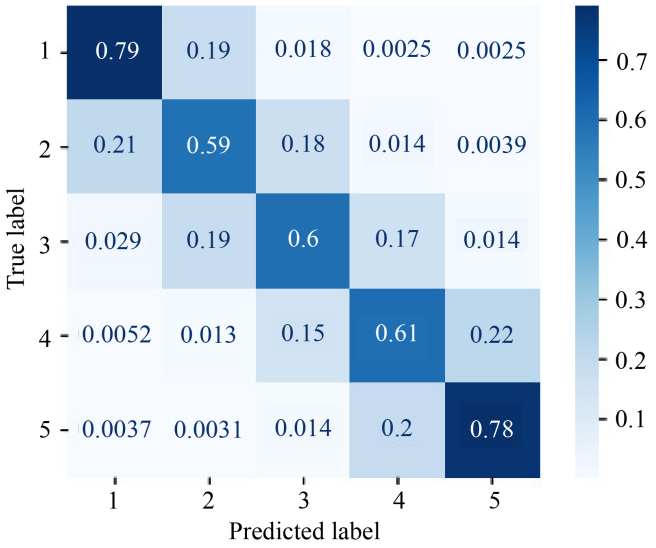

Figure 5. Confusion Matrix of Cased Base BERT on Test Set
图 5. 测试集上 Cased Base BERT 的混淆矩阵

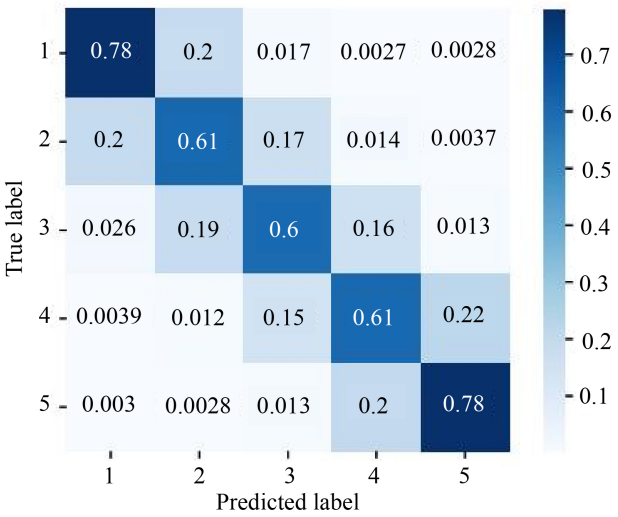

Figure 6. Confusion Matrix of Cased Large BERT on Test Set
图 6. 测试集上 Cased Large BERT 的混淆矩阵

4.3.2. DistilBERT

DistilBERT 模型是 BERT 模型的提炼，其中保留了 97%的 BERT 语言理解能力，提高了 60%的性能[17]。在测试集上的结果如表 4 所示，混淆矩阵如图 7 所示。可以发现，与多参数的 BERT 相比，DistilBERT 模型的精度降低了约为 1.36%，但在运行速度上比基础 BERT 模型快了一倍。当计算资源有限时，DistilBERT 将是一个很有吸引力的选择。

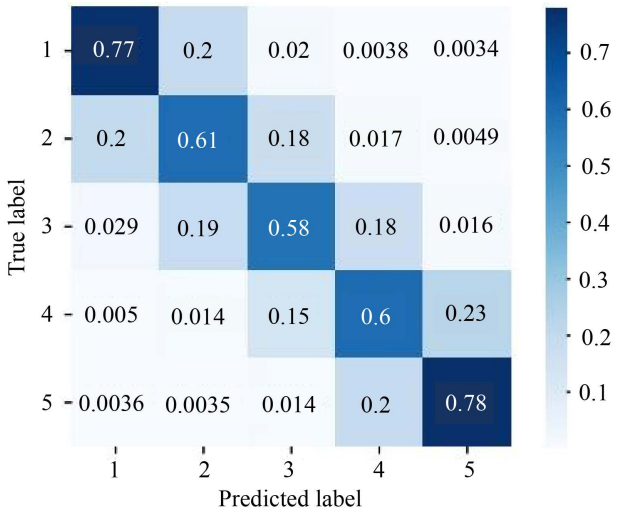

Figure 7. Confusion Matrix of DistilBERT on Test Set
图 7. 测试集上 DistilBERT 的混淆矩阵

4.3.3. RoBERTa

RoBERTa [16]改进了预训练程序，并在几个 NLP 任务上取得了先进的结果。本文实验中采用预训练的 roberta-base 架构进行实验，相关实验结果见表 4，混淆矩阵如图 8 所示。可以发现，该模型是最优的，与多参数的 BERT 模型相比，两者类似的运行时间的情况下，准确率提高了 0.62%，且有更高的得分。

这表明 RoBERTa 模型将是不考虑计算资源时更高预测指标的优选。

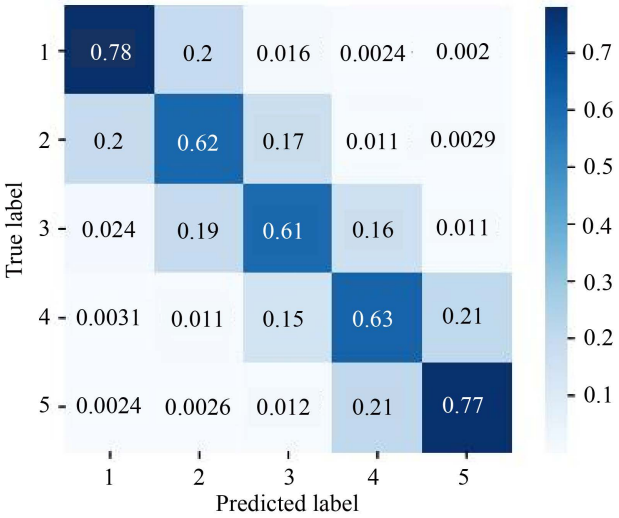

Figure 8. Confusion Matrix of RoBERTa on Test Set
图 8. 测试集上 RoBERTa 的混淆矩阵

5. 结论

在本文中，从 Yelp 大型数据集的评论文本中预测了评分。首先对 Yelp 开放数据集进行详尽的统计分析，阐述其分布概况，并创新性地建立了一个平衡的训练数据集，以适应之后的训练。其次基于 Tf-idf 向量器进行数值表示，采用了朴素贝叶斯和逻辑回归两种经典机器学习模型进行情感预测，与 BERT、DistilBERT 和 RoBERTa 三种基于 transformer 的模型比较，经过实验验证，发现以下两点结论：1) 在训练时间上机器学习方法实现轻松，运行时间短，可以作为基准，而较为先进的三种模型中，DistilBERT 模型训练时间最有优势，在计算资源有限的情况下是一个较佳的选择。2) 在分类效果上，多参数的 BERT 模型比基础 BERT 模型高出 1.02 个百分点，DistilBERT 模型的指标略低，而 RoBERTa 模型达到了 69.75% 的准确率，与该数据集上的最高水平相近。本文对 transformer 预训练模型进行了验证，未来将以此为基础，进行多模态模型的探索。此外，由于计算资源的限制，本文模型效果尚有提升空间。

参考文献

[1] Liu, S.Q. (2020) Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models. ArXiv: 2004.13851.

[2] 刘兵. 情感分析: 挖掘观点、情感和情绪[M]. 北京: 机械工业出版社, 2019: 149-156.

[3] Zhou, Z.Y. and Liu, F.A. (2021) Filter Gate Network Based on Multi-Head Attention for Aspect-Level Sentiment Classification. Neurocomputing, **441**, 214-225, https://doi.org/10.1016/j.neucom.2021.02.041

[4] Yang, Z.L., Dai, Z.H., Yang, Y.M., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. (2019) XLNet: Generalized Auto-regressive Pretraining for Language Understanding. ArXiv:1906.08237

[5] Guda, B.P.R., Garimella, A. and Chhaya, N. (2021) EmpathBERT: A Bert-Based Framework for Demographic-Aware Empathy Prediction. ArXiv Preprint ArXiv: 2102.00272.

[6] Yu, B.Y., Zhou, J.X., Zhang, Y. and Cao, Y.N. (2017) Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. ArXiv: 1709.08698.

[7] Asghar, N. (2016) Yelp Dataset Challenge: Review Rating Prediction. ArXiv: 1605.05362.

[8] Perez, L. (2017) Predicting Yelp Star Reviews Based on Network Structure with Deep Learning. ArXiv: 1712.04350.

[9] Yelp Open Dataset. https://www.yelp.com/dataset

[10] Cui, Y. (2015) An Evaluation of Yelp Dataset. ArXiv: 1512.06915.

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2001) The Elements of Statistical Learning. In: *Springer Series in Statistics*, Springer, New York. https://doi.org/10.1007/978-0-387-21606-5

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. ArXiv: 1706.03762.

[14] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ArXiv: 1810.04805.

[15] Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) Distilbert, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. ArXiv: 1910.01108.

[16] Liu, Y.H., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.Q., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) Roberta: A Robustly Optimized BERT Pretraining Approach. ArXiv: 1907.11692.

[17] Guda, B.P.R., Srivastava, M. and Karkhanis, D. (2022) Sentiment Analysis: Predicting Yelp Scores. ArXiv: 2201.07999.