多目标情感分类中文数据集构建及分析研究

刘鹏远　　田永胜　　杜成玉　　邱立坤
北京语言大学信息科学学院
国家语言资源监测与研究平面媒体中心
闽江学院计算机与控制工程学院
liupengyuan@pku.edu.cn　blcutys@gmail.com　du_chengyu@163.com　qiulikun@pku.edu.cn

摘要

目标级情感分类任务是要得到句子中特定评价目标的情感倾向。一个评论句中往往存在多个目标，多个目标的情感可能一致，也可能不一致。但在已有针对目标级情感分类的评测数据集中：1）大多数是一个句子一个目标；2）在少数有多个目标的句子中，多个目标情感倾向分布很不均衡，多个目标情感一致的情形占较大优势。数据集本身的缺陷限制了模型针对多个目标进行情感分类的提升空间。针对以上问题，本文构建了一个针对多目标情感分类的中文数据集，人工标注了6339个评价目标，共2071条数据。该数据集：1）评价目标个数分布平衡；2）情感正负极性分布平衡；3）多目标情感倾向分布平衡。随后，本文利用多个目标情感分类的主流模型在该数据集上进行了实验与比较分析。结果表明现有主流模型尚不能对存在多个目标且目标情感倾向性不一致实例中的目标进行很好的分类，尤其是目标的情感倾向为中性时。多目标情感分类任务具有一定的难度与挑战性。

关键词： 目标级情感分类；中文数据集；多目标

Construction and Analysis of Chinese Multi-Target Sentiment Classification Dataset

PengYuan Liu　　YongSheng Tian　　ChengYu Du　　Likun Qiu
Beijing Language and Culture University, School of Information Science
Language Resources Monitoring and Reserch Center
Minjiang University,School of Computer and Control Engineering
liupengyuan@pku.edu.cn　blcutys@gmail.com　du_chengyu@163.com　qiulikun@pku.edu.cn

Abstract

Target-level sentiment classification task is to get the sentiment tendency of a specific evaluation target in a sentence. There are often multiple targets in a comment sentence, and the sentiments of multiple targets may be consistent or inconsistent. However, in the existing evaluation datasets for target-level sentiment classification: 1) most of them are one sentence with one target; 2) in a few sentences with multiple targets, the sentiment distribution of multiple target is very unbanlance, and the situation where the sentiments of multiple targets are consistent has a great advantage. The defect of the dataset itself limits the improvement space of the model for sentiment classification for multiple targets. In response to the above problems, this paper constructs a Chinese dataset for multi-target sentiment classification, manually annotated 6339 targets, a total of 2071 items. The data set: 1) the distribution of the number of evaluation targets is balanced; 2) the distribution of positive and negative sentiments is balanced; 3) the distribution of multi-target sentimental tendency is balanced. Subsequently,

第十九届中国计算语言学大会论文集，第604页-第615页，海南，中国，2020年10月30日至11月1日。

this article uses multiple mainstream models of target-level sentiment classification to conduct experiments and comparative analysis on this dataset. Experimental results show that the existing mainstream models are still unable to well classify the targets in instances where there are multiple targets and the target's sentiment is inconsistent, especially when the target's sentiment is neutral. The task of multi-target sentiment classification is difficult and challenging.

Keywords: Target-level Sentiment Classification , Chinese Dataset , Multi-target

1 引言

社交网络、电子商务和网络新闻的发展迅速，每时每刻都有大量的评论和观点涌现。这些观点和评论文本中包含着非常重要的信息，比如通过分析某个商品的用户评论可以帮助潜在用户选择商品，也可以帮助企业改良商品等，因此情感分析成为自然语言处理领域中最活跃的研究问题之一。方面级别情感分析（Aspect-Level Sentiment Analysis）是一种细粒度的情感分析任务，关注文本针对某一实体、实体的某个部分或属性的情感倾向。目标情感分类（Aspect Term Polarity）是方面级别情感分析的核心子任务之一，目的是分析评论中的目标（Aspect Term）的情感倾向，这个目标是实体的一部分或者是实体的属性，且该目标必须明确出现在句子内。比如（见图1）：“大堂小了点儿，房间挺干净，价钱不错。”这句话中的目标词有“大堂”、“房间”和“价钱”，根据上下文“小了点”、“挺干净”及“不错”可以确定他们的情感倾向分别为负向、正向和正向。当前方面级别的情感分类的研究主要基于深度神经网络，采用端到端的方式进行情感倾向的预测或分类如(Dong et al., 2014; Vo and Zhang, 2015)。而循环神经网络因其在处理序列方面的优势得到了研究者更多的青睐，如(Tang et al., 2015; Ruder et al., 2016)等人的研究。除此之外，注意力机制(Cho et al., 2014)也常常被用来融合方面词与上下文的信息，(Wang et al., 2016a; Ma et al., 2017; Peng et al., 2017; 曾锋；曾碧卿；韩旭丽；张敏；商齐, 2019) 等利用方面词与句子进行交互以得到更好的表示。近年来，也出现了基于预训练语言模型BERT(Devlin et al., 2018)的方面级情感分类研究(Song et al., 2019; 杜成玉；刘鹏远, 2019)。

以上各类模型与方法通常会在同一个目标情感数据集上进行试验及横向比较，目前使用最广泛的目标情感分类数据集是SemEval-2014 task4(Pontiki et al., 2014)和Twitter(Jiang et al., 2011)，均为英文数据集。我们对这两个数据集中的评价目标和对应的情感倾向进行统计并发现：1）Twitter数据集及SemEval-2014 task4数据集中中含有一个以内评价目标的句子比例很高（分别为100%及73.6%）；2）SemEval-2014 task4数据集中，评价目标情感倾向不一致的句子占比仅有8.6%。而在实际应用中，一个句子包含一个以上目标词且评价倾向性不同的情况比较常见。但是，由于评测数据集中的多目标实例较少，情感不一致的实例更少，这种分布对现有模型在多目标句子上的目标情感分类评价造成一定困难，也限制了模型针对多个目标句子进行目标情感分类的提升空间。

为解决以上问题，本文构建了一个针对多目标情感分类的中文数据集[0]，人工标注了6339个评价目标，共2071条数据。该数据集：1）评价目标个数分布相对平衡；2）情感正负极分布相对平衡；3）多目标情感倾向分布相对平衡。随后，本文利用多个目标情感分类的主流模型在该数据集上进行了实验，比较了各个模型针对多目标情感分类的表现，并进行了详细分析与讨论。

2 数据集构建

2.1 数据准备

本文选取了谭松波收集整理的酒店评论语料料[1]作为原始语。该语料规模为10000篇，内

基金项目：北京市自然科学基金资助项目（4192057）资助

[0]https://github.com/NLPBLCU/Chinese-Multi-Target-Sentiment-Classification-Dataset
[1]https://languageresources.github.io/

```
<sentence id="2">
    <text>大堂小了点，房间挺干净，价钱不错。</text>
    <aspectTerms>
        <aspectTerm term="大堂" polarity="negative"  from="0" to="2"/>
        <aspectTerm term="房间" polarity="positive"  from="6" to="8"/>
        <aspectTerm term="价钱" polarity="positive"  from="12" to="14"/>
    </aspectTerms>
</sentence>
```

Figure 1: XML格式目标情感分类示例

容为携程网的评论。我们将原始语料进行去重并以此为待标注对象，共得到7767篇评论，其中5323篇是正向评论，2444篇是负向评论。

2.2 标注对象

依照目标情感分类任务研究的惯例，我们舍弃目标评价倾向为冲突的句子，仅标注为正向、负向与中性的评价句。

在语料选择方面，选取含有2个或以上目标的句子为标注对象，并且尽量控制标注数据在三种情况下的分布同时基本平衡：1）句中目标数量；2）目标情感正负倾向极性；3）句中目标情感相同或不同。

具体标注时，将每个评论中的目标词抽离出来，分别标注每一个目标词在其中的情感倾向，然后得到该句子的标注结果如："大堂"：负向；"房间"：正向；"价钱"：正向。

2.3 标注流程、数据格式与标注规范

由三名语言学及应用语言学专业的硕士生担任标注员进行标注，先进行一轮试标注，并进行讨论，在此基础上总结出标注规范。然后依据标注规范，由两名标注员独立地进行标注，对于标注不一致的情况由第三位标注员进行仲裁。

试标注。我们从待标注对象中随机抽取100条，按照标注程序进行试标注。三名标注员在标注了所有100条语料后，就标注不一致的数据进行讨论，总结并形成了标注规范[2]。

数据格式。数据用XML格式存储，如图1所示。其中：$< aspectTerms >$是目标及其情感倾向的标签，$< term >$表示目标，$< polarity >$表示情感倾向，$< from >$和$< to >$表示目标词开始位置和结束位置。

标注规范

（1）目标词：目标词是明确出现在句子中的被评价对象的具体属性，本节构建的是酒店领域的数据集，因此被评价对象是"酒店"，目标词可能是"装修风格"、"服务态度"等。只标注具有多个目标词的句子。

（2）情感倾向：包含正向、负向、中性三种情况。正向评价是评价者对某个目标词持积极的、满意的态度。负向评价是评价者对某个目标词持消极的、不满的态度。中性评价是评价者对某个目标词持中立的、客观的态度。

（3）标注单位：参照现有的英文目标情感分类数据集，本节以单个句子为单位进行标注。

（4）标注边界：只标注目标词，目标词前的形容词性修饰成分及数量短语等不在标注范围内。

（5）目标词包含名词型和动词型两种。

（6）目标词若出现多次，只标注离评价词最近的目标词。

2.4 标注结果

我们仅标注目标词为2的实例与目标词大于等于3的实例，最终标注了2071条数据，共6339个目标，平均每个句子3.06个目标。标注好的数据集基本情况如表1，2所示。句中目标词情感极性一致与不一致句子，目标词为正向情感与负向情感三者比例分别平衡。数据集整体标注一致率为78.1%。

[2]详细规范与示例将随数据集及代码一并发布。

目标词数量或情感倾向	数量	占比%
句子中目标词数量：2	954	46.1
句子中目标词数量：⩾3	1157	55.9
句子目标词情感极性一致	1009	48.7
句子目标词情感极性不一致	1062	51.3

Table 1: 目标词数量或情感倾向分布

	正向情感	负向情感	中性情感	总计
目标词数量	3001	2827	511	6339
目标词比例	47.3%	44.6%	8.1%	100%

Table 2: 所有目标词的情感倾向分布

3 模型与实验

3.1 模型

为探索和分析目标情感分类的主流方法在本文构建数据集上的表现，我们选择了5个具有代表性且已开源的主流神经网络模型，其中包括2个基于BERT的目标情感分类模型。我们还实现了1个基于BERT的基线模型BERT-SPC。

IAN(Ma et al., 2017):将上下文词与目标词通过LSTM层得到隐藏层状态序列，接着利用池化函数得到目标词的初始向量表示，该向量与上下文隐藏层状态通过注意力层得到上下文词注意力权重分布，接着计算加权后的上下文表示，将它最终的上下文向量。然后用类似的方法得到目标词表示，再与上下文表示拼接。

RAM(Peng et al., 2017)：首先通过双向LSTM层得到句子的隐藏层状态序列，接着利用位置信息构建位置加权记忆矩阵，然后构建多个注意力层，每一层的结果都是基于上一层的结果重新进行计算，以此来捕捉记忆矩阵中有用的信息。

ATAE-LSTM (Wang et al., 2016b)：对句子和给定的方面词用LSTM进行编码后，采用注意力机制对隐藏层输出进行处理，将得到的注意力向量与方面词向量拼接得到关于方面词的情感极性表达。

AEN-BERT(Song et al., 2019)：运用标签平滑化的方法来解决中性类别方面词情感模糊的问题，并运用了多个不同注意力机制对上下文和方面词进行建模。

BERT-HAN(杜成玉；刘鹏远, 2019)：建立于BERT上的基于螺旋注意力机制的神经网络模型。首先利用目标词构建句子，接着采用句子对的输入方式利用BERT预训练词向量，然后利用螺旋上下文注意力层和螺旋目标词注意力层通过多次叠加注意力层来更好地表示上下文和目标词。

BERT-SPC：使用预先训练好的BERT来生成序列的词向量，BERT有单个句子和句子对两种输入方式。本文采用句子对的输入方式，将目标词与上下文组成句子对进行输入，输入方式为“[CLS]+ target+[SEP]+context+[SEP]”，然后将得到向量送入softmax分类器。

3.2 实验

3.2.1 数据集与评价指标

数据集采用本文建立的多目标情感分类中文数据集，其中含有多个评价目标的句子共6339条。按照3:1的比例分别将两个数据集划分成训练集和测试集，具体划分见表3。评价指标采用分类准确率，即模型正确分类的样本数与模型总样本数之比。

数据集	正向目标数量/占比(%)	负面目标数量/占比(%)	中性目标数量/占比(%)	总计
训练集	2250/47.3	2120/44.5	383/8.1	4753
测试集	751/47.2	707/44.6	128/8.1	1586

Table 3: 多目标情感分类中文数据集详细信息

模型	词向量维度	隐状态维度	学习率	batch size
IAN	300	300	1e-3	16
RAM	300	300	1e-3	16
ATAE-LSTM	300	300	1e-3	16
AEN-BERT	768	N/A	2.00e-05	16
BERT-HAN	768	300	2.00e-05	16
BERT-SPC	768	N/A	2.00e-05	16

Table 4: 模型参数设置

	IAN	RAM	ATAE-lSTM	AEN-BERT	BERT-HAN	BERT-SPC
准确率(%)	74.1	82.7	79.3	75.1	81.5	**86.2**

Table 5: 模型的准确率

3.2.2 参数设置

表4是各模型实验时的参数设置情况。其中，IAN、RAM采用Stanford大学发布的GloVe词向量[3]来作为预训练词向量；BERT-SPC、AEN-BERT、BERT-HAN采用BERT BASE[4]进行预训练。

3.2.3 实验结果

实验结果如表5所示。在6个模型中，基线模型BERT-SPC表现最好，IAN表现最差。仅将目标词与实例一起输入并进行训练的BERT-SPC模型，就已经能够学到很好的目标词情感倾向信息。而其他两种进行目标词与句子进行不同注意力权重计算的模型：AEN-BERT与BERT-HAN的表现反而不如BERT-SPC。值得注意的是，两个非BERT模型RAM与ATAE-LSTM的表现分别比两个基于BERT的模型BERT-HAN与AEN-BERT要好，其可能的原因，我们将在后续分析中进一步尝试探究。

4 讨论

4.1 目标不同情感倾向对模型分类性能的影响

我们将所有模型对目标分别为正、负及中性三种情感倾向分类时的性能进行对比，结果如表6所示。其中正向及中性情感倾向最优模型为BERT-SPC，负向为BERT-HAN。所有模型在目标情感倾向不同时性能从好到坏的排序均为：目标为正向情感倾向>目标为负向情感倾向>目标为中性情感倾向。当目标情感为中性时，各个模型的表现均不尽人意。这一点很大程度上与数据集中目标情感倾向的分布有关（分布见本文第2小节中的表2）。

图1是将数据集中三种情感倾向目标的分布作为待比较的分布基准，考察所有模型性能提升绝对值的柱状图。可以看出，所有模型在目标为正/负向情感倾向时，性能较分布基准均有所提升，且提升幅度较大。目标为中性情感时，三种基于BERT的模型，不但性能较分布基准均有所提升且幅度较大（BERT-SPC提升幅度最大），这说明基于BERT的模型能在数据分布不均衡的条件下，学到一定的中性倾向的信息；而对其他三种非BERT模型，性能均低于分布基准，说明这几个模型所学到的目标中性情感倾向信息较少，甚至基本学不到（对IAN模型）。

基于BERT的模型能够比非BERT模型更好地融合中性情感目标数据的信息，我们推测可能会在一定程度上影响其在目标为正/负向情感时的性能，这可能是造成AEN-BERT与BERT-BERT在总体性能上没有RAM模型好的原因之一。

4.2 目标数对应模型分类性能的影响

为考察各模型在不同目标个数情况下的性能，根据数据集中每句含有的目标个数，分两类对各模型性能进行统计：1）含两个目标；2）含有三个目标及以上。

[3]https://nlp.stanford.edu/projects/glove/
[4]https://github.com/google-research/bert

模型	正向情感倾向	负向情感倾向	中性情感倾向
IAN	82.3	78.9	0.0
RAM	89.1	89.7	6.2
ATAE-LSTM	85.1	87.0	3.1
AEN-BERT	80.2	79.9	22.7
BERT-HAN	82.7	**92.1**	27.7
BERT-SPC	**91.3**	89.3	**39.1**

Table 6: 模型在不同情感倾向上的分类性能（准确率%）

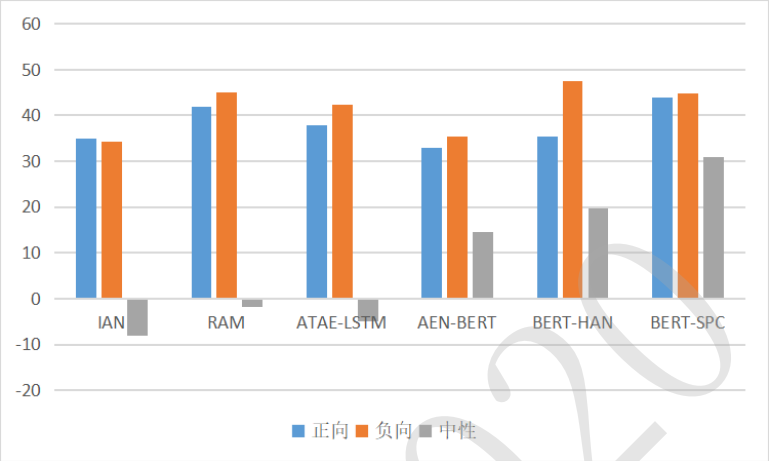

Figure 2: 各个模型分别在目标为三种情感倾向时性能较分布基准提升绝对值的柱状图（准确率%）

此外，我们还在本文第二节介绍的待标注对象语料中，按照本文建设多目标情感分类数据集基本类似的过程，额外标注了仅含有单个评价目标的例句共1046条，其情感倾向性分布与本文建立的多目标情感分类数据集基本一致。类似的，按照3:1的比例划分成训练集和测试集，分别为784条和262条。在这个单目标数据集上，所有模型重新进行了实验，参数设置与之前多目标实验相同（见表4）。

将各模型在单目标数据集及多目标数据集上的实验结果合并列于表7。其中，目标词数量为1，代表单目标数据集，其余两行代表在多目标数据集上的结果。

我们发现，虽然直觉上多目标数据相对于单目标数据更加难以分类，但在单目标数据集上的模型的性能并非均比在多目标数据集上的性能更高。同时，目标词数量分别在1,2及大于等于3时，模型的分类性能并没有较大的差距。

我们统计了在多目标数据集中，同一个条目多个目标情感一致与不一致时各个模型分类性能的表现，如表8所示。可以看出，在多个目标情感一致时，各个模型的性能均非常优异（大于90%），AEN-BERT的表现最好，其次是BERT-SPC。各个模型在多目标情感倾向不一致时，BERT-SPC的表现最好，能够达到80%，AEN-BERT表现最差，还不到60%。所有模型的性能比情感倾向一致时的性能均有大幅度的下降，降幅最大的是AEN-BERT，接近40%，降幅最小的是BERT-SPC，也达到了14.4%。

目标词数量	IAN	RAM	ATAE-lSTM	AEN-BERT	BERT-HAN	BERT-SPC
1	**81.1**	**83.7**	80.5	75.3	**82.1**	**89.3**
2	75.7	**83.7**	**80.9**	**79.0**	81.6	87.6
⩾3	73.4	82.7	79.6	73.84	81.7	85.6

Table 7: 不同目标数时模型的分类性能（准确率%）

多目标情感	IAN	RAM	ATAE-lSTM	AEN-BERT	BERT-HAN	BERT-SPC
一致	92.8	90.4	90.7	**98.1**	92.0	94.4
不一致	60.2	73.6	70.8	58.4	69.4	**80.0**
下降幅度	-32.6	-16.8	-19.9	**-39.7**	-22.6	-14.4

Table 8: 模型在多目标情感倾向一致和不一致数据上的分类性能（准确率%）

4.3 模型分类结果的相关性

图3是6个模型在本数据集上的分类结果相关性热力图，它反映的是各个模型在数据集上的预测结果的相关性。如果两个模型预测结果的相关性比较高，则一个模型预测正确时，另一个模型预测也较大可能是正确的；一个模型预测错误时，另一个模型也较大可能是预测错误的。

从图3可知，6个模型的分类结果相关性都比较高，基本都在80%以上，其中ATAE-LSTM与RAM的相关性及BERT-HAN与RAM的相关性相对较高，超过了85%。

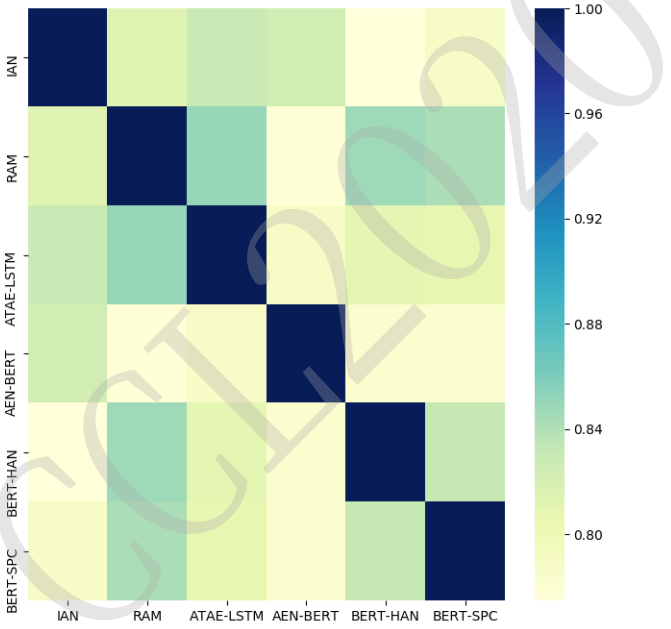

Figure 3: 模型分类结果的相关性

我们还统计了BERT-SPC模型预测错误（共219个）但其他模型能够预测正确的目标个数及比例（即与BERT-SPC模型预测错误目标个数之比），列在表9中。可见，虽然模型分类结果相关性比较高，但对BERT-SPC模型错误预测的目标，其他模型也有一定正确预测的可能，但是有部分目标（219-133=86个）所有模型均无法正确预测。

4.4 目标分类难度

数据集中目标的情感倾向可能被1或多个模型正确预测，也可能无法被任何模型正确预测。可从正确预测目标情感倾向模型个数的角度来考察目标情感倾向的预测难度。我们将各个模型对测试集中所有1586个目标实例的预测结果进行了逐个统计，对每一个目标进行模型预测正确计数，即每有一个模型对其预测正确，则该目标模型预测正确个数加一。所有目标分成从0到6共7类。在此基础上，我们进一步将所有目标分为易、中、难三个等级，结果列在表10：

易：模型预测正确个数为5及6的目标；

	IAN	RAM	ATAE-lSTM	AEN-BERT	BERT-HAN	总计
个数	61	89	84	78	86	133
比例%	27.6	40.6	38.4	35.6	39.3	60.7

Table 9: BERT-SPC模型预测错误但其他模型能够预测正确的目标个数及比例（与BERT-SPC模型预测错误目标个数之比）。其中总计是针对BERT-SPC模型预测错误而其他模型预测正确并去重后的目标总和及相应比例。

模型计数	0	1	2	3	4	5	6
目标个数	86	50	73	105	153	236	883
所占比例%	5.4	3.1	4.6	6.6	9.6	14.9	55.7
难度等级	难			中		易	
目标个数	136			331		1179	
所占比例%	8.6			20.9		70.5	

Table 10: 目标情感倾向预测难度等级分布

中：模型预测正确个数为2,3,4的目标；
难：模型预测正确个数为0,1的目标。

由表10可知，有86个目标，所有模型均没有预测正确，有883个目标所有模型均预测正确。难度等级为“难”、“中”及“易”的目标，分别占所有目标的8.6%，20.9%及70.5%。

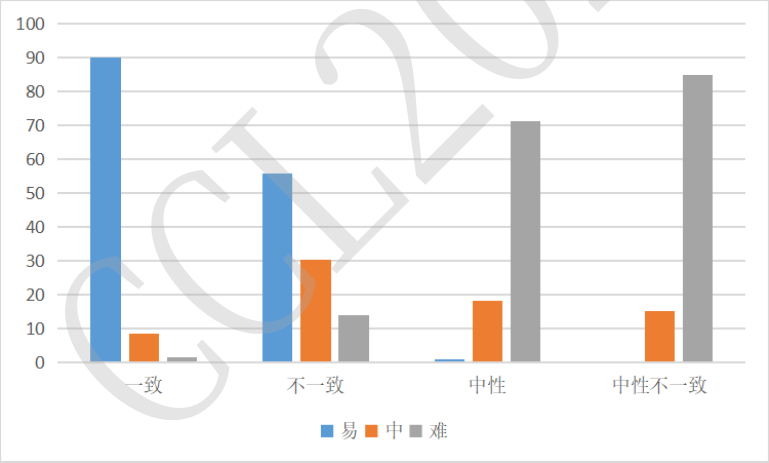

Figure 4: 目标“一致”、“不一致”，“中性”、“中性不一致”时的难度分布

在目标难度等级的基础上，我们考虑了以下四种典型情况：
1）**一致**：处在有情感倾向一致目标的句中的目标；
2）**不一致**：处在有情感倾向不一致目标的句中的目标；
3）**中性**：中性情感目标；
4）**中性不一致**：中性情感目标，且处在有情感倾向不一致目标的句中。

我们绘制了这四种典型情况下难度分布柱状图，见图4。由图4可知，当目标在“一致”时，90%的目标在等级“易”中，即容易分类，“不一致”时，仅有50%左右的目标容易分类。当目标为“中性”时，超过70%的目标在等级“难”中，即难以分类，而当目标为“中性不一致”时，有约85%的目标难以分类。这四种典型情况，目标分类难度从难到易依次为：“中性不一致”，“中性”，“不一致”，“一致”。

4.5 实例分析

为对模型难以分类的实例有一定感性认识，我们对全部模型均预测错误的目标例句进行

了仔细观察，并列举了部分在表11。表11中，其中，第一列为样例，红色字体为目标，红色字体后括号内的数字为该目标的情感倾向，1/-1/0分别表示正向/负向/中性；黑体词为本句要预测情感倾向的目标词。第二列为要预测情感倾向的目标，第三列为该目标类型，“一致”/“不一致”与4.4小节中考虑的前两种典型情况相同。第四列为要判断的目标词的真实情感倾向。最后一列是6个模型分别对目标词情感倾向的预测结果。

我们发现目标词以外的情感倾向会对目标词的情感倾向抽取造成干扰。如在句子“其实格林豪泰...还挺好，接着就是前台了，那叫什么态度啊？”中，目标词为“前台”，其情感极性很明显为负向。但是所有的模型都把他分到了正向，这可能是因为，在这个句子中短语“还挺好”离目标词更近，且情感倾向为正，但这是对目标词“大门”的描述，因此对模型的分类造成了干扰。在中性目标词的情感倾向分类上，由于目标词没有明显情感倾向，则其余目标词的情感倾向造成的干扰将会更加明显。此外，中性目标词的情感倾向含糊不清，程度较弱，不太容易判断他的情感倾向是否是中性。例如在句子“第二次入住了...冰镇饮料喝。”中，尽管目标词“冰镇饮料”的情感倾向为中性，但是句子中修饰目标词的“免费”很多情境下是表达正向的情感。在数据规模上，由于中性目标数量远远低于正向与负向目标数，故也使得模型对于中性目标的情感分类泛化能力较弱。

样例	目标	类型	情感	模型预测
其实格林豪泰给我的印象一直挺好的，（那是因为之前住的是上海的格林豪泰），所以就想换换环境，订了三天的房，首先一进大门（1），感觉还挺好，接着就是**前台**了，那叫什么服务态度（-1）呀?	前台	不一致	-1	1 1 1 1 1 1
但是在房间安排（-1）方面觉得有点欠妥，我订了两间高级房，一样的价格，但不一样的房型，一件[5]的漱洗间（-1）非常小，而且没有**阳台**，另一间的漱洗间却是这间的两倍，而且有阳台。	阳台	不一致	1	-1 -1 -1 -1 -1 -1
酒店早餐（-1）不是很丰富，房间设施（1）尚可，标准房无**矿泉水**送。	矿泉水	不一致	0	1 -1 1 1 -1 -1
第二次入住了，酒店有免费接机的服务（0），坐到车上后还有免费的**冰镇饮料**喝。	冰镇饮料	一致	0	1 1 1 1 1 1

Table 11: 全部模型均错误预测的部分实例。其中，第一列为样例，红色字体为目标，红色字体后括号内的数字为该目标的情感倾向，1/-1/0分别表示正向/负向/中性；黑体词为本句要预测情感倾向的目标词。第二列为要预测情感倾向的目标，第三列为该目标类型，“一致”/“不一致”与4.4小节中考虑的前两种典型情况相同。第四列为要判断的目标词的真实情感倾向。最后一列是6个模型分别对目标词情感倾向的预测结果。

5 相关工作

传统的基于方面词的情感分析方法包括基于规则的方法(Ding et al., 2008)和基于统计的方法(Jiang et al., 2011; Zhao et al., 2010)。这些方法侧重于将一组分类线索转化为特征向量，但这既需要费力的特征工程工作，也需要大量的额外语言资源。循环神经网络较早应用到方面级别情感分类领域(Tang et al., 2015; Ruder et al., 2016)。单纯基于RNN 的模型无法很好地捕捉到句子中与方面词与情感极性词或短语之间的关联，研究人员引入注意力机制来解决这个问题。Wang et al. (2016a)对句子和给定的方面词用LSTM进行编码后，采用注意力机制对隐藏层输出进行处理，得到关于方面词的情感极性表达。Tang et al. (2016)基于输入句子的词向量构成的外部记忆进行注意力学习，模型的每一层基于上一层输出的结果重新计算注意力分布，最终得到关于给定方面词的情感极性表达。Ma et al. (2017)不仅计算句子隐藏层输出的注意力分布，还计算方面词的注意力分布。Huang and Carley (2018)以联合的方式建模方面词和句子，

[5]因为是用户撰写的评论，因此存在一些用户自行输入的错误，为保证数据的真实性，本文未进行任何更正。

明确捕捉方面词和上下文句子之间的交互。Li et al. (2018)将位置嵌入作为输入的一部分，并用层次注意力机制来融合目标和上下文词的信息。卷积神经网络能够并行计算，在运算速度上有一定优势，于是也有学者基于参数化卷积神经网络(Huang and Carley, 2018)和基于门控卷积神经网络(Xue and Li, 2018)的相关研究。BERT提出后，一些研究在它基础上对上下文进行编码，并结合注意力机制，来更好地解决方面级别情感分类任务。Song et al. (2019)在BERT表示的基础上，采用多个不同注意力机制对上下文和方面词进行建模。类似的工作还有(Zhao et al., 2020)。杜成玉；刘鹏远 (2019)利用螺旋注意力机制，反复增强BERT编码后的方面词与句子的表示。

现有的方面级别情感分类任务的数据集主要有：SemEval-2014 task 4 Aspect Based Sentiment Analysis(Pontiki et al., 2014)、SemEval-2015 task 12 Aspect Based Sentiment Analysis(Pontiki et al., 2015)、SemEval-2016 task 5 Aspect Based Sentiment Analysis(Pontiki et al., 2016)和Twitter(Jiang et al., 2011)。

SemEval-2014 task4数据集包含两个领域，分别为laptop和restaurant。Laptop数据集摘自笔记本电脑的用户评论，包含3048个英语句子；Restaurant数据集来自于Ganu et al. (2009)标注的餐厅评论，由3044个英语句子组成。两个领域的数据集都以句子为单位人工标注了句子中的方面词及其情感倾向和位置，其中情感倾向包含正向、负向、中性，除此之外，Restaurant数据集还标注了方面词的类别的情感倾向。SemEval-2015 task 12数据集采用的原始语料与SemEval-2014 task4数据集相同，但它是以一条评论为单位进行标注的，两个领域的数据集都标注了方面词的类别及其情感倾向，但与SemEval-2014 task4数据集不同的是，它是实体和属性对，其中实体和属性属于提前规定好的实体和属性集合；Restaurant数据集还标注了观点目标词及其情感倾向及位置，它与SemEval-2014 task4数据集中的方面词的概念相同。SemEval-2016 task 5数据集的标注内容与SemEval-2015 task 12数据集相同，增加了其他语种的数据集，比如中文、法语、阿拉伯语等。Twitter数据集是使用关键字通过twitter API收集的开放域数据集，关键字包含名人、公司和产品的名称等，然后以tweet为标注单位，人工标注了tweet中出现的关键字及其情感倾向，这是目前为止人工标注的最大的针对方面情感分类任务的Twitter数据集。

6 结论

本文针对现有数据集的问题构建了一个针对多目标情感分类的中文数据集，该数据集评价目标个数、情感正负极性及多目标情感倾向均分布平衡。我们还实现了多个目标情感分类的主流模型并在该数据集上进行了实验与比较分析。结果表明：1）目标个数对各模型在数据集上的分类结果影响不大；2）在同一句中多个目标情感倾向是否一致对模型的影响较大；3）情感倾向为中性的实例较难进行预测，一方面是由于中性目标实例较少，另一方面是因为中性情感倾向的强度一般较低。多目标情感分类的模型应考虑如何对目标情感倾向性不一致，尤其是目标情感倾向性不一致且同时目标的情感倾向有中性情感的情况下，进行有针对性的改进。

致谢

本文受北京市自然科学基金资助项目（4192057）资助。感谢匿名评阅人的建议。

参考文献

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti, editors, *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 231–240. ACM.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland, June. Association for Computational Linguistics.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases, WebDB 2009, Providence, Rhode Island, USA, June 28, 2009*.

Binxuan Huang and Kathleen M. Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1091–1096. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 151–160. The Association for Computer Linguistics.

Lishuang Li, Yang Liu, and Anqiao Zhou. 2018. Hierarchical attention based position-aware network for aspect-level sentiment analysis. In Anna Korhonen and Ivan Titov, editors, *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 181–189. Association for Computational Linguistics.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Chen Peng, Zhongqian Sun, Lidong Bing, and Yang Wei. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 486–495. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *CoRR*, abs/1902.09314.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *Computer Science*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 214–224. The Association for Computational Linguistics.

Duy Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *International Conference on Artificial Intelligence*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Zhao Li. 2016a. Attention-based lstm for aspect-level sentiment classification. In *Conference on Empirical Methods in Natural Language Processing*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016b. Attention-based LSTM for aspect-level sentiment classification. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615. The Association for Computational Linguistics.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2514–2523. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 56–65. ACL.

Pinlong Zhao, Linlin Hou, and Ou Wu. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowl. Based Syst.*, 193:105443.

曾锋；曾碧卿；韩旭丽；张敏；商齐. 2019. 基于双层注意力循环神经网络的方面级情感分析. 中文信息学报, 33(6):108.

杜成玉；刘鹏远. 2019. 基于螺旋注意力网络的方面级别情感分析模型. In 第十八届全国计算语言学学术会议.