第 5 章 多模态多标签情感分析数据集构建

5.1 概　　述

现有的多模态情感数据集中均含有统一的多模态情感标注，没有独立的单模态情感标注。因此，本章将先构建一个多模态多标签的中文多模态情感分析数据集，对于每个多模态片段，同时包含一个多模态和 3 个单模态情感标签。然后，构建有监督的多任务多模态情感分析框架，在框架中引入 3 个主流的融合结构，通过对比实验充分验证单模态子任务对多模态主任务的辅助作用。

5.2 多模态多标签的中文情感分析数据集制作

本节构建了中文的多模态情感分析数据集（chinese single- and multi- modal sentiment analysis dataset，SIMS）。除了说话人语言上的差异，相比其他数据集，SIMS 数据集中除含有多模态情感标注外，还包含独立的单模态情感标注，如图 5.1 所示。在后续内容中将详细介绍此数据集的收集和标注过程。

图 5.1　SIMS 数据集与现有多模态数据集之间的差异

5.2.1 数据收集和标注

1. 数据收集

与单模态数据相比，多模态数据具有更高的收集要求。由于多模态情感分析主要研

究说话人的情感，因此，一个最基本的要求是说话人的脸部和声音必须同时在视频画面中出现并且持续一段时间。为了获取的视频片段尽可能接近日常生活，本章主要从电影、电视剧和生活类综艺节目中获取数据原材料。然后，结合视频剪辑工具 Adobe Premiere Pro[①] 对原素材进行帧级别剪辑，这是一个非常耗时但是足够准确的过程。此外，在收集过程中，以下三条准则被严格恪守。

（1）说话人语言为普通话，并且过滤掉带有地方口音的视频片段。

（2）视频片段的长度应在 1～10s。

（3）视频片段中有且仅有当前说话人的脸部出现。

最终，收集了 60 个原视频、2281 个有效视频片段。SIMS 具有丰富的人物背景，较大的年龄范围及高质量的数据内容，其详细的统计信息如表 5.1 所示[②]。之后，使用 FFmpeg 工具[③]从视频中分离出纯音频数据，再以人工方式对其进行语音转译获取对应的文本信息。

表 5.1　SIMS 数据集信息统计表

项　　目	数　量	项　　目	数　量
原视频总数	60	独立说话人总数	474
有效片段总数	2281	片段平均时长（秒）	3.67
男性	1500	片段中平均字数	15
女性	781		

2. 数据标注

在此部分，经过一定训练的 5 位独立标注者被邀请对每个视频片段进行多重情感标注。由于每个视频片段都需要包含一个多模态和 3 个单模态情感标注，因此，如何避免其他模态信息对当前待标注模态的信息干扰，是此过程着重考虑的一个问题。为了尽可能避免这种干扰现象，每位标注者被要求按照“文本→音频→无声视频→多模态”的顺序进行标注，并且在完成一种模态的情感标注后需要间隔一段时间才能进行另一种模态的标注。

然后，每位标注者给所有数据指定三分类情感标签：消极（－1）、中性（0）、积极（1）。与现有数据集[10,67]类似，为了使 SIMS 能够同时用于情感回归和分类任务，将 5 个标注值

① https://www.adobe.com/products/premiere.html。

② 已咨询过相关律师，仅用于学术目的的短视频数据集的制作和分发符合我国相关法律规定。

③ https://www.ffmpeg.org。

的均值作为最终的标注结果。于是，标注结果值在区间[−1,1]，分类和回归标签之间的对应关系如表 5.2 所示。

表 5.2　SIMS 数据集中分类标签和回归标签对应关系表

分类标签	回归标签	分类标签	回归标签
强消极情感	−1.0,−0.8	弱积极情感	0.2,0.4,0.6
弱消极情感	−0.6,−0.4,−0.2	强积极情感	0.8,1.0
中性情感	0.0		

5.2.2　统计和分析

首先，分析 SIMS 数据集不同模态中情感类别的分布倾向性，统计结果如图 5.2(a)所示。从图中可以看出，SIMS 数据集的情感更多地偏向消极，这可能是因为 SIMS 中的视频素材主要来自于电影等表演性影视作品，而这种作品中往往会有更多的消极表达，以此突出演员的表演能力。

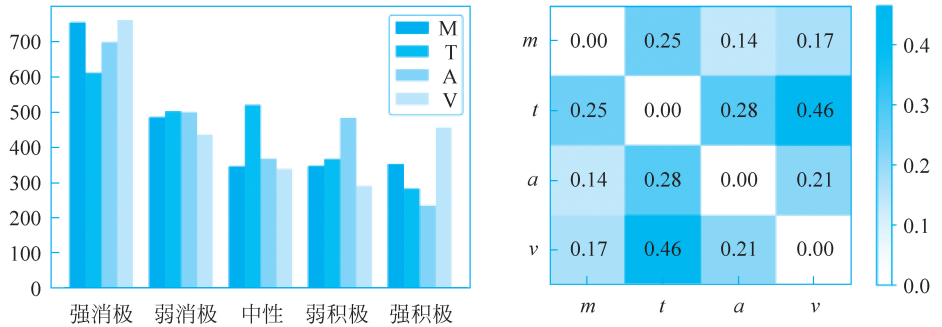

图 5.2　标注结果统计直方图和不同模态情感标签差异对比

其次，为了验证本章的初始动机——统一的多模态标签并不是时刻适用于单模态数据。此处绘制了不同模态情感标签之间的差异性混淆矩阵，如图 5.2(b)所示。图中的数值表示两个模态标签之间的差异性大小，值越大意味着情感差异性越大，其计算公式如下：

$$D_{ij}=\frac{1}{N}\sum_{n=1}^{N}(A_i^n-A_j^n)^2 \tag{5.1}$$

其中，$i,j\in\{m,t,a,v\}$；N 是样本数量；A_i^n 表示模态 i 中的第 n 个标签值。

从图 5.2 中可以看出，在音频和多模态之间的情感差异性最小(0.14)，而文本和视频之间的差异性最大(0.46)。这是因为音频信息中本身包含文本内容，更接近多模态信息，但是文本和无声视频之间并不存在直接联系。可见，图 5.2 得到的观察结果是符合经验预期的，也侧面印证了数据标注过程的可靠性。

至此，完成了 SIMS 数据集的构建工作，为后续工作奠定了数据基础。因此，第 6 章将基于此数据集构建多模态和单模态情感分析任务的联合学习模型，验证单模态子任务的引入是否能够辅助模型学到更有效的特征表示，进而提升多模态学习效果。

5.3　本 章 小 结

现有的多模态情感数据集中均仅含有统一的多模态情感标注，没有独立的单模态情感标注，并且缺少中文多模态情感分析数据集。因此，本章构建了一个多模态多标签的中文多模态情感分析数据集，对于每一个多模态片段，同时包含一个多模态和 3 个单模态情感标签；然后，构建有监督的多任务多模态情感分析框架，在框架中引入 3 个主流的融合结构，通过对比实验充分验证单模态子任务对多模态主任务的辅助作用，进一步验证构建多模态多标签的中文多模态情感分析数据集的有效性。