arXiv:2302.13321v1 [cs.SD] 26 Feb 2023

音乐中的多模态：基于高层次音频特征和歌词预测音乐情绪

Tibor Krols、Yana Nikolova 和 Ninell Oldenburg*

哥本哈根大学北欧研究与语言学系和
计算机科学系

摘要

本文旨在检验：在使用歌曲高层次特征和歌词时，多模态音乐情感识别（MER）方法是否优于单模态方法。我们从 Spotify 应用程序接口获取 11 项歌曲特征，并结合包含情感倾向、TF-IDF 和 Anew 的歌词特征，使用 4 种不同的回归模型，预测 Deezer 情绪检测数据集（DMDD）上的效价和唤醒度（Russell，1980年）得分（Delbouys 等人，2018年）。我们发现，在 11 项歌曲高层次特征中，主要是其中 5 项对模型表现有贡献；预测效价时，多模态特征的表现优于单独使用音频特征。我们已公开提供代码[1]。

1 引言

音乐的情感效应既是经过充分研究的事实（Hunter & Schellenberg，2010年；Balkwill & Thompson，1999 年；Deng 等，2015 年），也为大多数人所熟悉。此外，音乐情感识别中的跨文化一致性（Balkwill & Thompson，1999 年；Fritz 等，2009年）表明，基本的认知能力参与了这一过程。随着音乐数字化和情感计算技术的兴起，人们越来越关注从音乐中提取情感信息，以用于研究、治疗（Dingle 等，2015年；Bernatzky 等，2011年）或音乐推荐（Deng 等，2015年；Park 等，2015年）。这项研究属于音乐情感识别（Music Emotion Recognition），简称 MER（Chen 等，2015 年）。如今，甚至 Spotify 等音乐流媒体平台也配备了一个应用程序接口，使广泛的特征能够被访问，从而向公众开放。

但哪些特征能真正预测歌曲的情绪？Spotify 的标注表现如何？基于第2 节介绍的既有文献，我们假设结合高级听觉特征和从歌词中提取的特征的多模态方法，其表现优于单模态方法(Y.-H. Yang、Lin、Cheng 等人，2008年；Hu 和 Downie，2010b，2010a）。我们将在第3 节介绍音乐情感识别模型，随后在第4 节和第5节呈现并讨论探索性实验和回归实验的结果。

2 相关工作

随着数字音乐的兴起，利用音乐自动提取信息或进行预测的方法研究也随之展开。年度音乐信息检索评测交流会（MIREX）[2] 于 2007 年首次将音频音乐情绪纳入分类任务（MIREX，2007年），当时便已发现，真实标注和人类的情绪判断在这项任务中发挥着关键作用（Hu 等人，2008年）。所采用的特征范围已涵盖从频谱等低级特征（例如原始频率信号、频谱质心和标准差（X. Yang 等人，2018年））到“可舞性”（Hu 等人，2008年）、节奏或音色等高级特征（X. Yang 等人，2018年）。这些特征可从低级特征中提取，并被认为更贴近人类感知（Panda 等人，2021）。此后，相关系统也持续得到改进（Kim 等人，2010年）。

尽管文献普遍将音乐情感识别区分为三项任务：i）将情绪归入不同类别，ii）将情绪视为回归问题，以及 iii）检测歌曲中的情绪变化（Y.-H. Yang 和 Chen，2012年），我们将重点关注音乐情感识别中的第二项任务，即回归

*作者按字母顺序排列，且贡献相同

[1] https://github.com/Tibor-Krols/CogSci2-Spotify

[2] 一种“Downie，2006年基于社区的正式评测框架”

因为人们认为这种方法对所表示情绪的刻画更准确（Guan 等人，2012年）。为此，Russell (1980年)提出的二维情绪模型，以效价（从负面到正面）和唤醒度（从被动到主动）的数值表示情绪，是应用最广泛的衡量指标（Huang 等人，2016年；Trigeor- gis 等人，2016年；Wang 等人，2011年）。例如，愤怒是一种高唤醒度、低效价的情绪。本研究采用这一框架，因为它简单易用，尽管它无法涵盖所有情绪（Cowen 等人，2020年），因此，其他音乐情感识别研究（Hu 和 Downie，2010a；Hu 等人，2009年）采用了更多的情感类别。

2.1 将音乐情感识别视为回归任务

针对这一回归问题，最早的研究由 Y.-H. Yang、Lin、Su 和 Chen（2008年）开展，他们使用低层听觉特征和支持向量回归（SVR），获得了最高达 0.282 和 0.538 的效价—唤醒度（VA）R^2 得分。这与 Soleymani 等人（2014年）提出的 SVR 方法类似；该方法基于超分段特征，获得了 0.42 和 0.52 的 VA 得分。随后，Bai 等人（2016年）的随机森林回归器（RFR）基于 548 维低层和高层音频特征，获得了 0.293 和 0.625 的 VA 得分。其他方法还包括多元线性回归（MLR）（X. Yang 等人，2018年；Vatolkin 和 Nagathil，2019年），以及多层感知机（MLP）（Soleymani 等人，2014年；Bhattarai 和 Lee，2019年）。

深度学习（DL）模型在广泛应用中的兴起，以及这些模型的快速高效，也影响了音乐情感识别。其中最早的双模态 DL 分类模型之一见于（Jeon 等人，2017年），该模型采用端到端方法处理歌词和原始音频，无需进行任何特征工程，并在中层融合这两种模态。与当时的单模态方法相比，他们的模型表现更佳，准确率达到 80%。他们还发现，歌词比音频更能准确地预测情绪。Delbouys 等人（2018年）的一项研究考察了传统特征工程方法与深度学习的性能。在 Deezer 情绪检测数据集（DMDD）上，他们从一组效价—唤醒度相关词中提取低层听觉特征和词嵌入，并将这两类特征的中层融合（第二层或更后面的层）和后期融合（加权输出）与端到端深度学习方法进行比较。他们发现：i）深度学习中的中层融合与特征工程取得了相当的结果；ii）多模态中层融合优于两种单模态模型（音频与文本）。

2.2 歌词作为预测指标

D. Yang 和 Lee (2004年），他们通常被视为最早将音频与歌词结合用于音乐情感识别建模的研究者，采用词袋（词袋）方法，并结合 General Inquirer 的心理词汇标注来对歌词建模。

Hu 等人 (2009年）；Hu 和 Downie(2010a，2010b）；Hu 等人 (2017年）；Laurier 等人 (2008年）将歌词特征扩展到 n 元语法计数、风格特征和 ANEW 情感倾向标注（Bradley 和 Lang，1998年），并发现，在情感分类和回归任务中，歌词往往优于听觉特征，不过将这些特征结合起来才能取得最佳结果。如上所述，Delbouys 等人 (2018年）的深度学习模型在词嵌入方面也呈现出相同的规律。

在其近期研究中，Han、Kong、Han 和 Wang(2022年）报告称，词袋、TF-IDF 和词嵌入是音乐情感识别中最常用的文本表示方法。

2.3 高层特征

关于特征与这两种预测类别之间的关系，Gabrielsson 和 Lindstr¨om（2001）以及Juthiet al.（2020）指出，唤醒度与速度（快/慢）、音高（高/低）、响度水平（高/低）和音色（明亮/柔和）有关；效价则与调式（大调/小调）和和声（协和/不协和）有关。Vatolkin 和 Nagathil（2019年）则指出，节奏对于效价预测也很重要。

Panda 等人（2021）率先评估了 12 项高层 Spotify 特征与效价和唤醒度预测准确率之间的相关性。他们发现，只有三项特征（能量、效价和原声性）具有相关性。

2.4 Spotify

Spotify 每月拥有 4.22 亿活跃用户，是市场上最大的音乐流媒体服务（Ruby，2022）。它让用户能够轻松访问 70+百万首歌曲（Ruby，2022），并提供应用程序接口，可用于提取带有标注的音乐作品定制数据集[3]。这类标注的流程并未向公众公开

[3] https://developer.spotify.com/documentation/web-api/

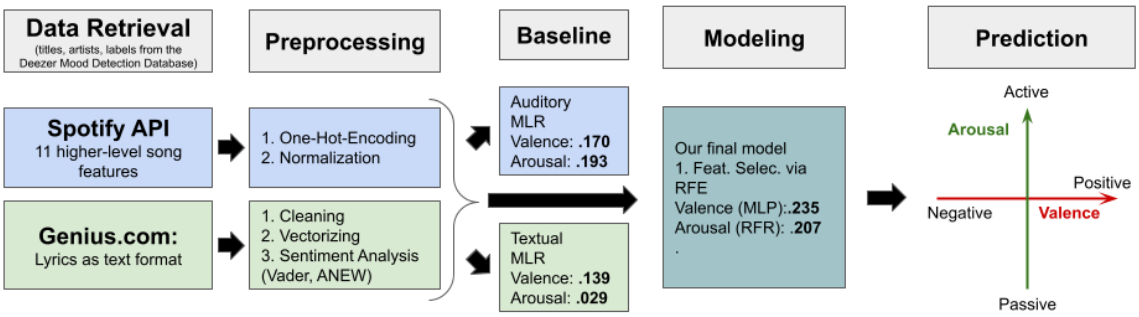


图 1：我们模型的流程。分数表示 R^2 在测试数据上的结果。

因为这是 Spotify 的主要收入来源之一（Hucker，2021）。不过，这种标注技术似乎是自动化的，由*TheEcho Nest*（2014年被 Spotify 收购）基于专家标注的数据开发。（*The Echo Nest*博客，2013）。

3 实验设置

3.1 数据

本研究使用的歌曲数据集是 Deezer 情绪检测数据集（DMDD）（Delbouys 等人，2018年），该数据集包含 18,644 首歌曲的效价-唤醒度分数，并基于百万歌曲数据集（Bertin-Mahieux、Ellis、Whitman 和 Lamere，2011 年）以及 LastFM 上与情绪相关的标签[4]（Hu 和 Downie，2010b）。效价-唤醒度分数是通过将扩展版 ANEW（XANEW）数据集应用于 LastFM 标签得到的。该数据集包含 14,000 个词及其相应的效价-唤醒度分数（Warriner 等人，2013；Bradley 和 Lang，1998）。

为了表示听觉模态，我们通过 Spotify API 从 DMDD 中获取了所有可用歌曲的听觉特征；这些特征被 X. Yang 等人（2018年描述为高层特征[5]，，最终获得 13,445 首歌曲，每首歌曲包含 11 个听觉特征。其中，原声性、可舞性、能量、器乐性、现场感、响度、语音性、速度和效价是连续变量，调式是二元变量，而调性是类别变量。我们为调性创建了虚拟变量，因此总计得到 23 个听觉特征。特征的详细说明见附录 A。Spotify 的效价标注与我们的真实效价采用不同的推导方式，从而避免循环论证；它也被用作情绪预测特征，见 Panda 等人（2021）。

最后，对于文本模态，我们抓取了 Genius.com 上可获取歌曲的歌词，这一做法由 Li、Mou 和 Chang（2018年）提出，从而将数据集进一步缩减至最终的 12,471 首歌曲，其中约 60% 用于训练，20% 用于验证，20% 用于测试。

3.1.1 提取歌词特征

为了表示歌词信息，我们创建了三类特征。

1. **情感信息**包括正向、负向、中性和复合分数，采用 Vader情感分析（Hutto & Gilbert，2014年）获得。

2.**TF-IDF特征**被广泛使用（Han 等人，2022），旨在提取歌曲文本中与效价和唤醒度相关的一般词汇特征。歌词文本经过词形还原后，获得 TF-IDF 单词元计数，并通过主成分分析降至 100 个主成分。

3.**扩展版 ANEW（XANEW）特征**旨在提取效价和唤醒度的词汇基础。该数据集在3.1中介绍。我们为每条预处理后的歌词文本生成两个计数向量，并将计数乘以相应的效价-唤醒度分数。这些计数特征也通过主成分分析降至 100 维。最终模型未纳入这些特征（见3.2）。

3.2 特征选择

基于这些特征，我们为两种不同的模态选择了特征子集（见图 2）。首先，我们在多元线性回归（MLR）中检验了显著性（与我们的基线类似，见第 3.3.1节），以获得听觉特征中的最优组合，最终得到五个显著预测变量：可舞性、能量、器乐性、效价和调式。

[4] https://www.last.fm/
[5] https://developer.spotify.com/console/

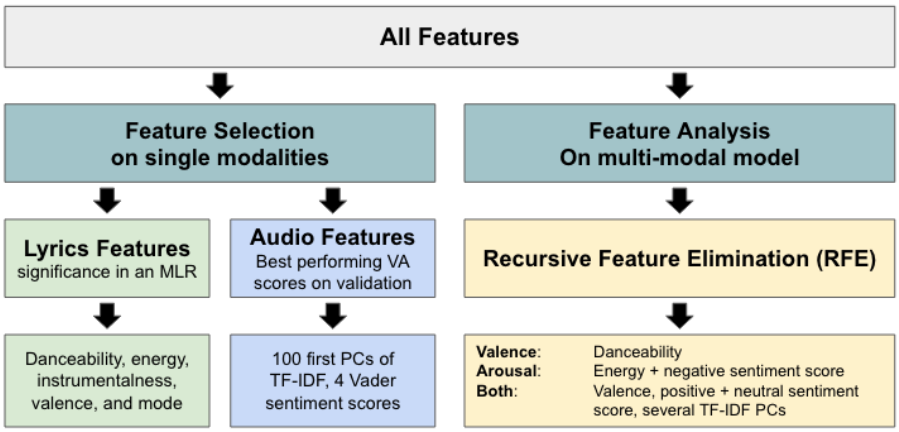


图 2：特征选择与分析流程

为了选择歌词特征，我们计算了由 Vader 情感、XANEW 和 TF-IDF类型特征构成的所有组合的效价-唤醒度分数（见附录 B）。Vader 与 TF-IDF特征的前100个主成分的组合在各回归模型中总体表现最佳，因此被用于我们的最终建模。

遵循 Hu 和 Downie(2010a、2010b）以及特别是Panda 等人 (2021），为确保结果可比，并旨在构建真正的多模态特征集，我们在特征层面合并了所有数据，形成一个数据矩阵 $M \in \mathbb{R}^{Nxd}$，包含 N =12,471 和 $d = 23 + 4 + 100 = 127$。

3.3 模型

基于第 2 节介绍的文献，我们比较了不同的回归模型，以评估基于高层听觉特征和歌词的单模态与多模态音乐情感识别方法的性能。所使用的回归学习算法均来自 `scikit-learn` 软件包（Pedregosa 等人，2011年）：线性回归，随机森林回归，

支持向量回归，
多层感知机回归器，
均使用我们通过网格搜索交叉验证获得的最优参数进行训练。

3.3.1 基线

我们为效价-唤醒度分数的单模态和多模态预测分别建立了多元线性回归（MLR）基线。对于音频基线，我们首先将所有听觉特征作为输入。其次，我们仅使用统计上显著的（$p < 0.05$）系数进行多元线性回归。然后，对于歌词基线，我们使用选定的歌词特征（情感倾向和 TF-IDF）进行多元线性回归。最后，我们将显著的音频特征与这些选定的歌词特征结合起来进行多元线性回归，构成多模态基线。

3.4 事后特征分析

此外，为了考察整体上最重要的一组特征，我们在多元线性回归模型上使用Granitto 等人，2006年提出的递归特征消除（RFE）方法，采用 `sklearn.feature-selection.RFE`软件包（Pedregosa 等人，2011年）。此外，我们还创建了一个包含所有音频特征的多元线性回归模型 + ，并纳入复合情感分数，以便分析系数（见表2）。

4 结果

我们按照上述文献（例如 R^2 中采用的方法，评估不同回归模型的Hu 等人(2008年，2009年）；Kim 等人(2010年)的预测能力。

4.1 模型结果

表 1总结了使用不同模态的回归模型的测试结果。对于效价和唤醒度，多模态特征均取得了总体最高的分数。对于效价而言，这一点更加明显：与仅音频特征和仅歌词特征相比，多模态显著提升了 R^2 。然而，在单模态模型中，音频特征对效价的预测效果略优于歌词特征。

唤醒度结果则更为复杂。尽管多模态随机森林回归取得了最高的唤醒度得分，但其他所有回归算法

Mode	模型	效价	唤醒度
音频	MLR	0.170	0.193
	RFR	0.171	**0.204**
	SVR	0.165	0.203
	MLP	**0.176**	0.203
歌词	MLR	**0.139**	**0.029**
	RFR	0.121	0.027
	SVR	0.042	-0.074
	MLP	0.117	0.020
多模态	MLR	**0.236**	0.190
	RFR	0.224	**0.207**
	SVR	0.208	0.154
	MLP	0.235	0.196

表 1： R^2 根据 4.2.1 中所述的特征子集，所有单模态和多模态模型的测试分数。

特征	效价	唤醒度
常数	**-1.6885***	**-0.9836***
可舞性	**0.6915***	**-0.3266***
能量	**0.6378***	**1.4254***
响度	-0.0091	-0.0073
语音性	-0.1101	**0.3952***
原声性	**0.1649***	0.0207
器乐性	0.0929	**-0.3278***
现场感	**0.1916***	0.0207
效价	**1.0901***	**0.5158***
速度	0.0005	0.0004
Mode	**0.0977***	**0.1272***
复合情感	**0.2275***	-0.0051

表 2：MLR 的系数。*表示显著，p <0.05

在使用单模态（音频）特征进行唤醒度预测时表现更好。这表明，加入歌词特征可能会产生干扰，并加剧唤醒度预测的误差，因为我们还发现，仅使用歌词时，唤醒度得分实际上为 0。

没有哪一种回归算法表现最佳，尽管支持向量回归的表现始终最差。这可能是由于超参数优化不足，因为针对这类模型测试更多参数组合非常耗时。

令人意外的是，对于仅使用歌词的效价和唤醒度，以及多模态效价，我们较复杂的回归算法并未优于基线多元线性回归。

4.2 特征分析

根据递归特征消除的结果，两个目标变量最重要的特征均包括效价、正向和中性情感分数，以及 TF-IDF 分析得出的若干主成分（见图 2 ）；此外，效价的相关特征还包括可舞蹈性，唤醒度的相关特征则包括能量和负向情感分数。

根据多元线性回归中系数的 p 值，效价有 7 个显著预测变量，而唤醒度有 6 个，如表 2 所示。

4.2.1 所选特征性能

表3 比较了多层感知机在各模态下使用全部特征与所选特征时的效价-唤醒度分数。特征子集如下：

1. 所有特征$_A$ = {原声性、可舞性、能量、器乐性、调性、现场感、响度、调式、语音性、速度、效价}2. 已选特征$_A$ = {可舞性、能量、器乐性、效价、调式}3. 所有特征$_L$ = {200ANEW 分数的主成分、TF-IDF 的 100 个主成分、4 个 VADER 情感分数}4. 已选特征$_L$ = {100 TF-IDF 的主成分、4 个 VADER 情感分数}

我们选择使用多层感知机来展示这些结果，因为它的性能稳定，且优化和训练所需时间较短。我们发现，使用选定的特征可使所有模态的 R^2 略有但稳定地提升，因此我们使用这些特征训练了最终模型。

总体而言，我们可以从表3 中看出，与使用全部特征相比，所有选定特征子集的表现始终更好。

	特征集	效价	唤醒度
音频	所有特征$_A$	0.163	0.193
	已选$_A$	**0.176**	**0.203**
歌词	所有特征$_L$	0.091	0.009
	已选$_L$	**0.117**	0.019
多模态	所有特征$_A$ + 所有特征$_L$	0.230	0.193
	选定$_A$ + 选定$_L$	**0.235**	**0.196**

表3：不同特征子集的多层感知机 R^2-分数比较

5 讨论

我们的主要结果可以简要概括为两点。第一，单模态歌词特征和单模态音频特征都能较为合理地预测效价，尽管多模态方法的表现优于单独使用任一模态。第二，使用歌词特征预测唤醒度较为困难，因为仅使用音频特征的表现几乎与多模态方法一样好。

5.1 特征的作用

效价。 如表 2所示，可舞性、能量、效价（Spotify）、调式、原声性、现场感和复合情感均对效价得分有正向贡献，其中前三项是最强的预测变量，优势明显。

唤醒度。 能量、语音性、调式和效价这些特征有正向贡献，而可舞性和器乐性的系数为负。能量是最强的预测变量，优势明显。

这两项发现与 Panda 等人 (2021)关于能量、效价和原声性的结论一致。然而，与他们不同的是，我们发现可舞性、器乐性、调式和语音性也与音乐情感识别相关。尤其是关于可舞性重要性的发现差异颇为有趣。在我们的基线模型中，可舞性的系数位居最高之列，而 Panda 等人 (2021)发现可舞性的特征权重最低。

根据 Spotify 的描述，可舞性反映了歌曲的速度、节奏和节拍特征，而感知响度和音色等感知特征则影响能量（见附录 A）。可舞性是效价的强预测变量，这支持了先前关于节奏属性对效价预测的重要性的发现（Y.-H.Yang, Lin, Su, & Chen, 2008年; Va-tolkin & Nagathil, 2019年）。我们的研究发现，能量是唤醒度的强预测变量，这间接支持了音色与唤醒度相关的研究（Gabrielsson & Lindstr¨om, 2001； Y.-H. Yang, Lin,Su,&Chen, 2008）。由于能量也反映感知响度，我们的结果同样支持响度与唤醒度之间的关联（Gabrielsson & Lindstr¨om, 2001； Y.-H. Yang, Lin, Su, &Chen, 2008）。

Spotify 的效价是效价最强预测指标这一发现，表明两者确实在某种程度上由同一潜在构念衡量，尽管其构建方式不同——。此外，我们发现调式与预测效价之间存在正相关系数，这与传统观点相符：大调与积极情绪相关，小调则与消极情绪相关（Gabrielsson& Lindstr¨om， 2001； Panda,Malheiro,& Paiva， 2020年）。

用于预测效价的情感分数正系数可以解释为：歌词更积极的歌曲具有更高的效价分数，而歌词更消极的歌曲具有更低的效价分数。

令人惊讶的是，根据 RFE 和 VA 验证分数，XANEW 特征属于预测能力最弱的特征之列，因此从最终模型中删除了这些特征。真实效价-唤醒度评分基于使用 XANEW 进行的标签分析（Delbouys 等人， 2018年），而 Hu 和 Downie （2010a）使用 ANEW 特征的研究取得了良好结果。XANEW 特征与 VA 分数之间的主要区别在于，前者基于 LastFM 上的标签，而后者基于歌词。这表明，歌曲歌词中的词语与人们在 LastFM 上为歌曲添加的标签存在 substantial 差异。TF-IDF 可能更擅长找出具有区分度的词语，因为它考虑了完整的歌词词汇。

5.2 多模态与单模态

效价。 我们关于模态的结果总体上与既有文献（Y.-H. Yang、Lin、Cheng 等人， 2008年; Hu 和 Downie， 2010a; Delbouys 等人， 2018年）基本一致，即多模态音乐情感识别模型优于单模态模型。然而，就我们的情况而言，这一点只有在效价方面得到了明确证实：歌词特征与音频特征结合后似乎产生了互补作用。根据单模态得分，音频对效价的贡献似乎更大；这一点与Ali 和 Peynircio˘glu （2006年）、Delbouys 等人（2018年）和Xue 等人 （2015年）的研究，以及轶事证据（Heinrich， 2019年）相一致，即人们对音乐的关注多于歌词；这会影响听众标注，而我们的目标变量正是基于这些标注得出的。另一方面，较早的研究，如Hu 和 Downie（2010a， 2010b），确实报告称，在效价预测方面，歌词可以优于音频。然而，这些研究采用了更广泛的特征工程和更细粒度的情绪类别（在二元分类任务中进行预测），因此限制了可比性。

唤醒度。关于唤醒度预测，多模态优势并不明确。我们的四个回归模型中，只有一个在采用多模态时表现更好，另外三个仅使用音频特征时表现更佳。歌词特征似乎要么没有贡献，要么会主动干扰音频特征，导致多模态表现下降。这与Delbouys 等人（2018年）的发现一致：唤醒度主要由节奏和速度等音频特征决定。Malheiro 等人（2018年）仅根据歌词就成功对歌曲的唤醒度和效价进行分类，但其数据集较小，且采用了非常广泛的特征工程，尤其纳入了更多风格和结构特征；而在我们的词袋 TF-IDF 方法中，这些特征已在预处理阶段被去除。另一方面，Delbouys 等人（2018年）没有报告对歌词进行预处理，而是将歌词直接输入多种神经网络；他们在同一数据集上得到的纯歌词模型的唤醒度得分同样较低。

5.3 总结与未来方向

最后，我们希望阐明模型表现的可能原因及其局限性，并展望未来研究方向。

早期特征融合。如3.2所述，我们选择在特征层面结合特征，遵循 Hu 和 Downie(2010a、2010b) 以及 Panda 等人(2021) 的做法。然而，Y.-H. Yang、Lin、Cheng 等人 (2008) 指出，早期融合的一个潜在难点在于实现真正共同的表示。尤其是将 TF-IDF 向量这类截然不同的特征类型与听觉特征结合起来，可能是导致我们结果的原因之一。对于该数据集，未来可以进一步研究后期融合方法（例如参见 Y.-H. Yang、Lin、Cheng 等人(2008)；Delbouys 等人 (2018)，这项探索性研究留待未来开展。未来研究的另一个有趣方向，是考察我们所使用特征之间的交互效应。

作为最先进水平的深度学习。有趣的是，我们发现结果与Delbouys 等人（2018年）在同一数据集上对比传统特征工程方法与深度学习模型的研究结果高度一致。使用多层感知机进行仅音频效价预测时，我们取得了有竞争力的分数（0.176），与他们的卷积网络（0.179）相当；而我们的多模态多元线性回归（0.236）甚至优于他们的中期融合深度学习网络（0.219），在效价方面。这表明，使用更高层次的特征能让我们的模型更具可解释性，同时不牺牲性能。

模糊的标注标准。我们得出的结果与 Delbouys 等人 (2018年) 相似，这表明部分局限可能与数据集本身有关。Spotify 的效价与预测效价相关性不高（R^2 = 0.35），这也意味着准确测量和评估可能较为困难。Delbouys 等人(2018年) 通过 LastFM 标签获得的效价—唤醒度标注未经人工标注核查或验证，因此其质量尚不明确。此外，目标效价—唤醒度变量假定整首歌曲只有一个数值，而没有呈现乐曲中效价和唤醒度可能出现的变化；然而，X. Yang 等人(2018年) 指出，歌曲的情感倾向可能会随着乐曲进程而变化，这属于音乐情感识别的子领域——音乐情感变化检测（MEVD）（Han 等人，2022）。

音频与预测。尽管纳入选定的 Spotify 特征后，我们发现总体得分有所提升，但其他直接取用歌曲频谱特征的方法（例如 Hu 等人 (2008年）；Hu 和 Downie(2010a）；Soleymani 等人 (2014年）至少在效价得分方面似乎表现更好，早在 2014 年就达到了 0.42。这与 Panda 等人 (2021) 的观点一致：Spotify 特征仍有改进和进一步明确的空间，而且仅凭高层特征仍然较难预测效价和唤醒度。

歌词与预测。早在首个音乐情绪分类任务中，MIREX(2007)便发现，音乐情感识别在很大程度上取决于真实标签的标注和人类判断，包括人们对歌词的关注程度。OutKast 的《Hey-Ya》便是一个例子：这首歌听起来欢快而充满活力（在我们的数据集中对应较高的效价和唤醒度得分），但艺术家本人却赋予它更为负面的含义（Archive-Corey-Moss，2004）。因此，无论对人类评判者还是预测模型而言，都很难确定这类歌曲真正的多模态效价。歌词篇幅短，以及隐喻、矛盾修辞和反讽等修辞手法的运用，更增加了难度，使歌词真正的情感倾向更难捕捉。我们建议未来采用更先进的情感分析方法。

6 结论

我们已经证明，可以结合音频和文本模态，为大型歌曲数据集预测效价和唤醒度；与单模态方法相比，这尤其提升了效价预测效果。使用我们所采用的特征，从歌词预测唤醒度的效果并不理想，可能需要更全面的特征工程或深度学习方法。总体而言，在根据能量、语音特征和调式等听觉特征预测唤醒度时，纯音频模型优于多模态模型。

参考文献

Ali, S. O., &Peynircio˘glu, Z. F. (2006年，10月)。歌曲与情感：歌词和旋律是平等的合作伙伴吗？音乐心理学，*34*(4), 511–534。检索日期：2022-05-30，来源：https://doi.org/10.1177/0305735606067168 (出版商：SAGE Publications Ltd) doi: 10 .1177/0305735606067168 6Archive-Corey-Moss. (2004年)。通往格莱美之路：*Outkast*《*Hey Ya!*》的制作过程。检索日期：2022-05-28，来源：https://www.mtv.com/news/1484730/road-to-the-grammys-the-making-of-outkasts-hey-ya/ 7Bai, J., Peng, J., Shi, J., Tang, D., Wu, Y., Li, J., & Luo, K. (2016年，8月)。通过效价—唤醒度回归进行维度音乐情绪识别。载于*2016*年 *IEEE* 第*15*届认知信息学与认知计算国际会议 *(ICCI*CC)*(第42–49页)。doi: 10.1109/ICCI-CC.2016.7862063 2Balkwill, L.-L., & Thompson, W. F. (1999年，10月)。音乐情绪感知的跨文化研究：心理物理与文化线索。音乐感知，*17*(1), 43–64。检索日期：2022-05-10，来源：https://online.ucpress.edu/mp/article/17/1/43/62058/A-Cross-Cultural-Investigation-of-the-Perception (出版商：加利福尼亚大学出版社) doi: 10.2307/40285811 1Bernatzky, G., Presch, M., Anderson, M., & Panksepp, J. (2011年，10月)。音乐作为一种非药物疼痛管理工具的情绪基础，载于 现代医学。《神经科学与生物行为评论》，35(9)，1989–1999。检索日期：2022-05-10，来源：https://www . sciencedirect.com/science/article/pii/S0149763411001151 doi: 10.1016/j.neubiorev.2011.06.005 1 Bertin-Mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011年1月). 百万歌曲数据集。载于（第591–596页）。3 Bhattarai, B., & Lee, J. (2019年6月). 使用迁移学习和多层感知机自动检测音乐情绪。《国际模糊逻辑与智能系统期刊》，19，88–96。doi: 10.5391/IJFIS.2019.19.2.88 2 Bradley, M. M., & Lang, P. J. (1998). 英语词汇情感规范（ANEW）：指导手册与情感评分。载于（第49页）。2, 3 Chen, S.-H., Lee, Y.-S., Hsieh, W.-C., & Wang, J.-C. (2015年12月). 使用深度高斯过程进行音乐情感识别。载于2015年亚太信号与信息处理协会年度峰会暨会议（APSIPA）（第495–498页）。doi: 10.1109/APSIPA.2015.7415321 1 Cowen, A., Fang, X., Sauter, D., & Keltner, D. (2020年). 音乐让我们感受到什么：至少有13个维度构成了不同文化中与音乐相关的主观体验。《美国国家科学院院刊》，117(4), 1924–1934。检索日期：2022-05-11，来源：https://www.pnas.org/doi/10.1073/pnas.1910704117 doi: 10 . 1073/pnas.1910704117 2 Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M. (2018年9月). 基于音频与歌词的音乐情绪检测：深度神经网络。arXiv:1809.*07276* *[*cs, stat*]*。检索日期：2022-05-03，来源：http://arxiv.org/abs/1809.07276（arXiv: 1809.07276）1, 2, 3, 6, 7 Deng, J. J., Leung, C. H. C., Milani, A., & Chen, L. (2015年3月). 与音乐相关的情绪状态：分类、变化预测及其在推荐中的考量。《ACM交互式智能系统汇刊》，5(1)，1–36。检索日期：2022-05-10，

来源：https://dl.acm.org/doi/10 . 1145/2723575 doi: 10.1145/2723575 1 Dingle, G. A., Kelly, P. J., Flynn, L. M., & Baker, F. A. (2015年9月). 音乐对成瘾治疗服务对象情绪和渴求的影响：一项针对两个临床样本的研究。《艺术与心理治疗》，45, 18–25。检索日期：2022-05-10，来源：https://www .sciencedirect.com/science/article/pii/S019745561500043X doi: 10.1016/j.aip.2015.05.005 1 Downie, J. S. (2006年12月). 音乐信息检索评估交流（MIREX）。《D-Lib杂志》，12(12)。检索日期：2022-05-31，来源：http://www.dlib.org/dlib/december06/downie/12downie.html doi: 10.1045/december2006-downie 1 The Echo Nest博客. (2013年11月). 检索日期：2022-05-31，来源：https://web.archive.org/web/20170422195736/http://blog.echonest.com/post/66097438564/plotting-musics -emotional-valence-1950-2013 3 Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., . . . Koelsch, S. (2009年4月). 音乐中三种基本情绪的普遍识别。《当代生物学》，19(7)，573–576。检索日期：2022-05-10，来源：https://www . sciencedirect.com/science/article/pii/S0960982209008136 doi: 10.1016/j.cub.2009.02.058 1 Gabrielsson, A., & Lindstr¨om, E. (2001). 音乐结构对情绪表达的影响。载于《音乐与情感：理论与研究》（第223–248页）。美国纽约：牛津大学出版社。2, 6 Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). 使用随机森林进行递归特征消除，以分析农工业产品的PTR-MS数据。载于（第8页）。4 Guan, D., Chen, X., & Yang, D. (2012). 基于多模态特征的音乐情绪回归。载于《国际计算机音乐建模与检索研讨会论文集》，70–77. 2 Han, D., Kong, Y., Han, J., & Wang, G. (2022年12月). 音乐情感识别综述。《计算机科学前沿》，16(6), 166335。检索于2022年5月18日，来源：https://link.springer.com/10.1007/s11704-021-0569-4 doi: 10.1007/s11704-021-0569-4 2, 3, 7 Heinrich, S. (2019年10月). 22首“有趣”的歌曲，其实歌词相当黑暗。检索于2022年6月1日，来源：https://www.buzzfeed.com/shelbyheinrich/catchy-songs -dark-lyrics（栏目：回顾）6 Hu, X., Choi, K., & Downie, J. S. (2017). 多模态音乐情绪分类的评估框架。《信息科学与技术协会会刊》，68(2), 273–285。doi: 10.1002/asi .23649 2 Hu, X., & Downie, J. S. (2010a). 通过结合歌词与音频，改进数字图书馆中的音乐情绪分类。载于《第十届数字图书馆年度联合会议论文集》（第159–168页）。检索于2022年5月4日，来源：https://doi.org/10.1145/1816123.1816146 1, 2, 4, 6, 7 Hu, X., & Downie, J. S. (2010b年8月). 当歌词在音乐情绪分类中胜过音频：特征分析。ISMIR，619–624。1, 2, 3, 4, 6, 7 Hu, X., Downie, J. S., & Ehmann, A. F. (2009). 音乐情绪分类中的歌词文本挖掘。海报展示，6。2, 4 Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). 2007年MIREX音频情绪分类任务：经验教训。载于（第6页）。1, 4, 7 Huang, M., Rong, W., Arjannikov, T., Jiang, N., & Xiong, Z. (2016). 基于双模态深度玻尔兹曼机的音乐情感分类。载于 A. E. Villa, P. Masulli, & A. J. Pons Rivero (编)，《人工神经网络与机器学习 – ICANN 2016》（第199–207页）。Cham：施普林格国际出版。doi: 10.1007/978-3 -319-44781-0 24 2 Hucker, M. (2021年11月). 揭秘Spotify算法的工作原理。检索于2022年5月31日，来源：https://

towardsdatascience.com/uncovering-how-the-spotify-algorithm-works-4d3c021ebc83Hunter, P. G., & Schellenberg, E. G. (2010). 音乐与情感。载于 M. Riess Jones, R. R. Fay, & A. N. Popper (编)，音乐感知（第129–164页）。纽约：施普林格。检索于2022年5月11日，来源：https://doi.org/10.1007/978-1-4419-6114-3_5 4doi:10.1007/978-1-4419-6114-3 5 1Hutto, C., & Gilbert, E. (2014年5月). VADER：一种简约的、基于规则的社交媒体文本情感分析模型。国际 *AAAI* 网络与社交媒体会议论文集，*8*(1), 216–225。检索于2022年5月10日，来源：https://ojs.aaai.org/index.php/ICWSM/article/view/14550（期号：1）3Jeon, B., Kim, C., Kim, A., Kim, D., Park, J., & Ha, J.-W. (2017). 基于端到端多模态神经网络的音乐情感识别。载于（卷：RecSys 海报，第2页）。2Juthi, J. H., Gomes, A., Bhuiyan, T., & Mah- mud, I. (2020). 运用机器学习方法提取音频特征进行音乐情感识别。载于 P. K. Singh, B. K. Panigrahi, N. K. Suryade- vara, S. K. Sharma, & A. P. Singh (编), *ICETIT 2019*论文集（第318–329页）。Cham：施普林格国际出版。doi: 10.1007/978-3-030-30577-2 27 2Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., . . . Turnbull, D. (2010). 音乐情感识别：前沿综述。载于（第12页）。1, 4Laurier, C., Grivolla, J., & Herrera, P. (2008年12月). 利用音频与歌词进行多模态音乐情绪分类。载于*2008*年第七届国际机器学习与应用会议（第688–693页）。doi: 10.1109/ICMLA.2008.96 2Li, S., Mou, C., & Chang, C. (2018). 基于歌曲歌词预测音乐流派与情绪。深度学习，斯坦福大学，加利福尼亚州，6。3Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2018年4月). 用于分类与回归的情绪相关特征音乐歌词。《IEEE情感计算汇刊》，9(2)，240–254。（会议名称：IEEE情感计算汇刊）doi：10.1109/TAFFC.2016 .2598569 7 MIREX。（2007）。2007：主页 - MIREX维基。检索于2022-05-31，网址：https://www.music-ir.org/mirex/wiki/2007:Main Page 1, 7 Panda, R., Malheiro, R. M., & Paiva, R. P.（2020）。用于音乐情感识别的音频特征：综述。《IEEE情感计算汇刊》，1–1。（会议名称：IEEE情感计算汇刊）doi：10.1109/TAFFC.2020.3032373 6 Panda, R., Redinho, H., Gonc˛alves, C., Malheiro, R., & Paiva, R. P.（2021 年 7 月）。Spotify 应用程序接口与音乐情感识别领域最先进方法相比如何？载于《第18届声音与音乐计算会议论文集》（SMC 2021）（第238–245页）。Axea sas/SMC Network。检索于2022-05-26，网址：https://estudogeral.sib . uc.pt/handle/10316/95161（接受日期：2021-07-04T16:31:54Z）doi：10 . 5281/zenodo.5045100 1, 2, 3, 4, 6, 7 Park, S.-H., Ihm, S.-Y., Jang, W.-I., Nasridinov, A., & Park, Y.-H.（2015）。一种利用排序属性进行情感识别的音乐推荐方法。载于 J. J. J. H. Park, I. Stojmenovic, H. Y. Jeong, & G. Yi（编），《计算机科学及其应用》（第1065–1070页）。柏林，海德堡：施普林格。doi：10.1007/978-3-662 -45402-2 151 1 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E.（2011）。Scikit-learn：Python中的机器学习。《机器学习研究期刊》，12(85)，2825–2830。检索于2022-05-26，网址：http://jmlr.org/papers/v12/pedregosa11a.html 4 Ruby, D.（2022年5月）。Spotify 2022年统计数据 —（事实、数据、信息图表）。检索于2022-05-31，网址：https://www.demandsage.com/spotify-stats/ 2 Russell, J. A.（1980年）。情感环形模型。《人格与社会期刊》

《心理学》，39(6)，1161–1178。（地点：美国；出版商：美国心理学会）doi：10.1037/h0077714 1, 2 Soleymani, M., Aljanaki, A., Yang, Y.-H., Caro, M. N., Eyben, F., Markov, K., . . . Wiering, F.（2014年11月）。音乐情感分析：方法比较。载于《第22届ACM国际多媒体会议论文集》（第1161–1164页）。美国纽约：计算机协会。检索于2022-05-26，网址：https://doi.org/10.1145/2647868.2655019 doi：10 . 1145/2647868.2655019 2, 7 Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S.（2016年3月）。特征再见？使用深度卷积循环网络进行端到端语音情感识别。载于《2016年IEEE声学、语音与信号处理国际会议》（ICASSP）（第5200–5204页）。（ISSN：2379-190X）doi：10.1109/ICASSP . 2016.7472669 2 Vatolkin, I., & Nagathil, A.（2019）。评估用于预测音乐唤醒度和效价的音频特征组。载于 N. Bauer, K. Ickstadt, K. L¨ubke, G. Szepannek, H. Trautmann, & M. Vichi（编），《统计计算的应用：从音乐数据分析到工业质量改进》（第305–326页）。Cham：施普林格国际出版。检索于2022-04-26，网址：https://doi.org/10.1007/978-3-030-25147-5 19 doi：10.1007/978-3-030-25147-5 19 2, 6 Wang, X., Chen, X., Yang, D., & Wu, Y.（2011）。基于歌词、使用Tf*idf和押韵对中文歌曲进行音乐情感分类。ISMIR。2 Warriner, A. B., Kuperman, V., & Brysbaert, M.（2013年12月）。13,915个英语词元的效价、唤醒度和支配度规范。《行为研究方法》，45(4)，1191–1207。检索于2022-05-10，网址：https://doi.org/10.3758/s13428-012-0314-x doi：10.3758/s13428-012-0314-x 3 Xue, H., Xue, L., & Su, F.（2015）。通过融合音频与歌词进行多模态音乐情绪分类。载于 X. He, S. Luo, D. Tao、C. Xu、J. Yang 和 M. A. Hasan（编），《多媒体建模》（第 26–37 页）。Cham：施普林格国际出版。doi：10.1007/978-3-319-14442-9 3 6 Yang, D., 和 Lee, W.-S.（2004）。使用软件代理消歧音乐情绪。2 Yang, X., Dong, Y., 和 Li, J.（2018年7月）。基于数据特征的音乐情绪识别方法综述。《多媒体系统》，24(4)，365–389。检索于 2022-05-10，网址：https://link.springer.com/article/10.1007/s00530-017-0559-4（公司：施普林格 分销商：施普林格 机构：施普林格 标签：施普林格 编号：4 出版商：施普林格柏林海德堡）doi：10.1007/s00530-017-0559-4 1, 2, 3, 7 Yang, Y.-H., 和 Chen, H. H.（2012年5月）。音乐情绪的机器识别：综述。《ACM智能系统与技术汇刊》，3(3)，40:1–40:30。检索于 2022-05-10，网址：https://doi.org/ 10.1145/2168752.2168754 doi：10.1145/2168752.2168754 1 Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., 和 Chen, H. H.（2008年）。迈向多模态音乐情绪分类。载于 Y.-M. R. Huang 等人（编），《多媒体信息处理进展——PCM 2008》（第 70–79 页）。柏林，海德堡：施普林格。doi：10.1007/978-3-540 - 89796-5 8 1, 6, 7 Yang, Y.-H., Lin, Y.-C., S u, Y.-F., 和 Chen, H. H.（2008年）。音乐情绪识别的回归方法。《IEEE音频、语音与语言处理汇刊》。doi：10.1109/TASL.2007.911513 2, 6

特征描述

原声性	衡量音轨为原声的置信度，取值范围为 0.0 至 1.0。1.0 表示置信度很高 该音轨为原声。 0 <= x <= 1
可舞性	根据多种音乐元素的综合情况，描述音轨适合跳舞的程度，包括 速度、节奏稳定性、节拍强度和整体规律性。0.0 表示最不适合跳舞，1.0 表示最适合跳舞。
能量	能量是一个介于 0.0 到 1.0 之间的数值，用于表示强度和活跃度的感知度量。 通常，充满活力的音轨听起来速度快、音量大且嘈杂。例如，死亡金属的能量很高，而 巴赫的前奏曲在该音阶上的得分较低。构成这一属性的感知特征包括动态 范围、感知响度、音色、起音率和总体熵。
器乐性	用于预测音轨是否不含人声。在此定义中，“呜呜”和“啊啊”等声音被视为器乐 语境。说唱或口语音轨显然属于“人声”。器乐性值越接近 1.0， 音轨不含人声内容的可能性就越大。高于 0.5 的值旨在表示 器乐音轨，但该值越接近 1.0，可信度越高。
Key	音轨所处的调性。整数使用标准音级记法映射到音高。例如 0 = C音、 1 = C/D♯、 2 = D音，依此类推。如果未检测到调性，该值为 -1。-1 <= x <= 11
现场感	检测录音中是否存在观众。现场感值越高，表示该录音由现场演出的可能性越大 ，表明该音轨为现场演出的可能性越高。值高于 0.8 时，强烈表明该 音轨为现场录音。
响度	音轨的整体响度，以分贝（dB）为单位。响度值是对整条音轨取平均后得到的 数值，可用于比较不同音轨的相对响度。响度是声音的一种属性，也是与物理强度（振幅） 最主要的心理学相关指标。其数值通常介于 -60 和 0 dB 之间。
Mode	调式表示音轨的调式（大调或小调），也就是其旋律所依据的音阶类型 内容由此得出。大调用 1 表示，小调用 0 表示。
语音性	语音性用于检测音轨中是否包含口语。音轨越接近纯口语， 录音（如脱口秀、有声书、诗歌），该属性值就越接近 1.0。高于 0.66 的数值 表示音轨很可能完全由口语构成。介于 0.33 和 0.66 之间的数值表示 音轨可能同时包含音乐和语音，可能分段呈现或彼此叠加，包括说唱等情况 音乐。数值低于 0.33 时，最有可能代表音乐及其他非语音类音轨。
速度	音轨估算出的总体速度，以每分钟节拍数（BPM）表示。在音乐术语中，速度是 一段乐曲的快慢，并直接取决于平均节拍时长。
效价	介于 0.0 到 1.0 之间的数值，用于描述音轨所传达的积极程度。高值音轨 效价较高的声音听起来更积极（例如开心、愉快、兴高采烈），而效价较低的音轨听起来则更 消极（例如悲伤、沮丧、愤怒）。 0 <= x <= 1

表4：所用特征及其描述摘自Spotify 文档。

B 歌词特征表现

	线性回归		RF		SVR		MLP	
特征	效价	唤醒度	效价	唤醒度	效价	唤醒度	效价	唤醒度
词频-逆文档频率	0.0705	0.0188	0.0631	0.0251	-0.0138	-0.0099	-0.8320	-2.0411
anew	-0.0146	-0.0138	0.0060	-0.0091	-0.0255	-0.0336	0.0121	-0.0026
VADER情感分析工具	0.0478	-0.0025	-0.0049	-0.0765	0.0076	-0.0003	0.0536	0.0041
词频-逆文档频率+ANEW	-0.0028	-0.0291	0.0427	-7.0790e-06	0.0765	-0.0056	**0.0962**	0.0167
词频-逆文档频率+VADER情感分析工具	**0.0931**	**0.0159**	**0.0842**	**0.0267**	0.0817	**0.0044**	0.0921	**0.0216**
ANEW+VADER情感分析工具	0.0238	-0.0187	0.0536	-0.0056	0.0284	-0.0223	0.0518	-0.0038
词频-逆文档频率+ANEW+VADER情感分析工具	0.0304	-0.0251	0.0796	0.0019	**0.0890**	0.0041	0.0911	0.0089

表5：不同歌词特征组合在效价和唤醒度上的验证 R^2 得分。词频-逆文档频率和 Vader 情感是效果最佳的特征组合。回归模型均采用其默认参数运行。