音乐中歌词和音频的联合情感分析

Lea Schaab [1] , Anna Kruspe [2]
[1] Technische Hochschule Nürnberg Georg Simon Ohm, Email: schaable79981@th-nuernberg.de
[2] Munich University of Applied Sciences, Email: anna.kruspe@hm.edu

1
赋予动机

音乐通常被描述为情感的语言，许多研究证实，听众将音乐视为情感 [?] 表达。情绪和情绪调节的表达，以及个人的自我发现和社会联系，是人们听音乐的主要原因之一 [?] 。国际唱片业联合会的“2022 年音乐参与”报告显示，69 名% 受访者认为音乐对他们的心理健康 [?] 很重要。

随着音乐数字可用性的快速增长，允许根据特定因素搜索和组织音乐的自动音乐信息检索系统的相关性也越来越高。虽然传统的音乐管理方法依赖于歌曲名称、艺术家或专辑名称等元数据，但它们通常对大多数与音乐相关的查询 [?] 适用性有限。近年来，基于感知到的情绪检索音乐变得越来越重要，尽管它仍处于早期发展阶段。这开辟了各种潜在的应用，包括音乐推荐系统、音乐搜索、音乐可视化、自动音乐生成、音乐治疗，或为各种环境（如餐厅、广告和电影制作 [?, ?] ）选择合适的背景音乐。

由于这些原因，音乐情感识别（MER）是音乐信息检索（MIR）社区中的一个热门研究课题。通常，情感仅根据歌曲的听觉内容来提取。相比之下，文本歌词在许多 MIR 任务中是一种研究不足的模式 [?, ?, ?] 。然而，歌词在引起听众的情绪反应方面起着重要作用 [?, ?] 。因此，在本文中，我们将研究听觉层面以及相应的文本歌词的情感识别。我们将将现有的机器学习模型用于两种不同的模式，比较结果，然后分析融合这些模型的选项。

Table 1: MIREX Mood 集群 [?]

Cluster	Mood
Cluster 1	passionate, rousing, confident, boisterous, rowdy
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good-natured
Cluster 3	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral

2
理论基础 情绪是心理学的一个关键研究课题，导致了音乐情感分析中对情绪进行分类的两种主要方法的发展：分类法和维度法。

分类方法根据基本情绪理论将情绪分为不同的类别，建议有限数量的主要情绪，如快乐、悲伤、愤怒、恐惧和厌恶，所有其他情绪都从中衍生出来 [?] 。这种方法已被应用于音乐信息检索交换（MIREX）音频情绪分类任务等研究和框架中，该任务将情绪分为五个集群，并带有相关的形容词 [?] 如表 ?? 所示。然而，它的缺点包括与人类感知的音乐情感范围相比，分辨率有限，并且由于语言描述 [?] 容易产生歧义。

另一方面，维度方法将情绪视为多维空间中的点，主要使用罗素的情感 [?] 环形模型。该模型利用两个维度：效价（愉悦）和唤醒（强度）来映射情绪，提供了一种简单而有效的方法来组织和比较不同的情绪，如图 ?? 所示。尽管有其优点，但维度方法可能会模糊重要的心理区别，并且可能无法仅通过几个维度 [?] 完全捕捉所有情绪。一些研究人员建议添加第三个维度，如支配，以提供更全面的情感表征，尽管这增加了复杂性，并且没有被普遍接受。

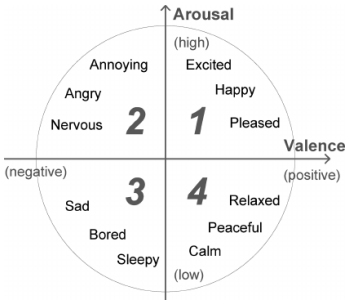

Figure 1: 二维效价-唤醒情绪空间 [?]

3
数据 数据集的选择基于音频和文本同时可用性、数据集大小、注释质量以及与已建立情感模型的兼容性等标准。从 MOODetector 项目 [?] 中选择了两个最符合这些标准的数据集。

VA 数据集
该数据集于 2016 年首次在 [?] 上提出，包括 133 个音频摘录及其相应的歌词，根据 Russell 的环形模型（即效价和唤醒值）进行注释。注释者被要求单独标记歌词和音乐。

类似 MIREX 的数据集
此数据集从 2013 年开始包含 903 个音频片段、764 个歌词和 193 个 MIDI 文件，元数据组织在 CSV 文件中，注释基于 All Music Guide [?] .

数据预处理
为了标准化这两个数据集，VA 数据集的文本注释根据其标记值与价唤醒象限对齐。对于类似 MIREX 的数据集，注释被分类到相应的象限中，利用从 ANEW 词典 [?] 得出的效价和唤醒值。映射可以在 `https://github.com/annakaa/sentiment_mappings` 下找到。其他预处理步骤包括使用歌词、出版年份和流派扩充数据集，规范化文本和音频，以及纠正不一致和重复。

4
音频模型 对于音频模态，我们采用了 USC SAIL 的模型，该模型在 2020 年的“音乐中的情感和主题” MediaEval 任务中表现最佳，该任务涉及预测音频中的标签以识别

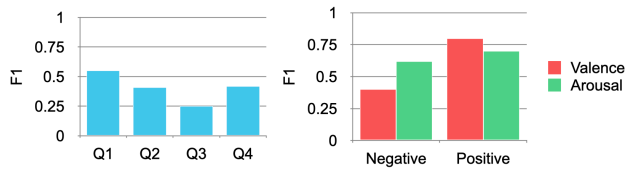

Figure 2: 纯音频模型的结果。左：环形模型的象限，右：效价和唤醒的二元结果。

与音乐相关的情绪和主题 [?] 。USC SAIL 模型采用具有残差连接的短块 CNN，并在包括 MTG-Jamendo 和 MagnaTagATune [?] 在内的数据集上进行训练。

为了使 MediaEval 标签适应数据集中组织的 Russell 环形模型的四个象限，遵循了前面为 MIREX 数据集描述的过程。根据 ANEW 词典，根据效价和唤醒维度将情绪相关标签分配给象限。映射也可在 `https://github.com/annakaa/sentiment_mappings` 下使用。

在改编过程中，人们注意到“爱”和“性感”这两个标签经常出现在所有象限中，导致它们被排除在进一步考虑之外，并被重新分配给主题标签。标签“史诗”和“沉重”主要与象限 Q2 相关，而 ANEW 词典中没有的“冥想”则根据个人判断主观分配给 Q4。

5

文本模型 评估了 Hugging Face 平台的四个模型，用于歌词的情感分析：finetuning-sentiment-model-4500-lyrics 该模型是专门为歌词中的情感预测而设计的，基于微调的 distilBERT 模型。它将歌词分类为负面或正面。[?] sentiment-roberta-large-english （SiEBERT）RoBERTa-large 的微调版本，针对性能进行了优化。SiEBERT 提供二进制预测，并在 15 个数据集中进行训练和评估。[?] bert-base-uncased-poems-sentiment 旨在将诗歌分为几类（消极、积极、中性、混合），探索该模型对歌词诗歌语言的适应性。[?] bert-base-uncased_finetuned_sentiments 经过训练，可以将文本分为六种不同的情绪（愤怒、恐惧、喜悦、爱、悲伤、惊讶），与二元情绪模型相比，提供多情绪视角。[?]

一个重大挑战是 512 个代币的最大代币数量限制，这是 BERT 模型的典型特征。为了管理超过此限制的歌词，文本被分割成块，单独处理，然后汇总以进行整体情绪预测。这种分块过程允许对较长的文本进行情感分析，从而扩大了分析范围。

6

实验结果

6

仅音频 南加州大学 SAIL 模型在音频情感分析中的表现在 Russell 的环形模型象限以及效价和唤醒维度上进行了评估。结果如图 ?? 所示。

对于罗素环形模型象限的分类，象限 Q1 显示出最好的结果，而象限 Q3 的性能指标最低。在价维度上，该模型在识别积极情绪方面表现出明显的优势，积极和消极情绪的平均指标约为 60 % 。该模型在唤醒维度上的表现平均高于价维度。在音频模式中，唤醒可能更容易确定，因为平静与兴奋的音乐表达更清晰。

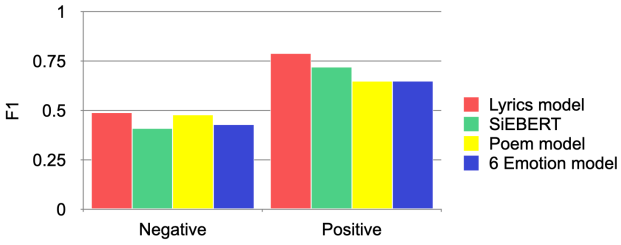

Figure 3: 各种纯文本模型的结果。

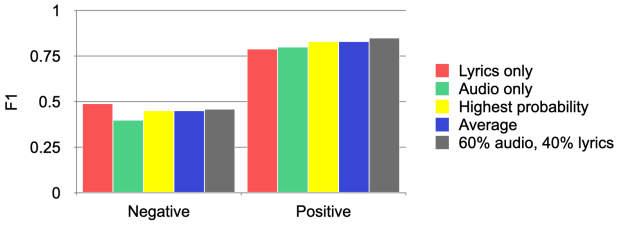

Figure 4: 不同模型融合策略的结果。

6

纯文本 对这四个模型进行了歌词情感分析的评估，由于模型之间的可用性一致，因此仅关注效价。结果如图 ?? 所示。歌词模型在检测积极情绪方面表现出色，对积极情绪的准确率和 F-Score 表现出更高的水平，表明对积极情绪的识别能力很强。SiEBERT 具有更广泛的训练基础，在两种情感类型上都略微落后于歌词模型，这表明像歌词模型这样的专业模型在歌词等特定任务中表现更好。

针对诗歌内容设计的诗歌模型在积极情绪分类方面表现出很高的精确性，在某些方面甚至优于歌词模型，凸显了其诗意性质对抒情分析的适用性。6 情绪模型涵盖了更广泛的情绪范围，表现出混合的表现，对负面情绪的回忆更好，但精度较低，突出了其识别负面情绪的能力，但偶尔会将积极情绪错误地归类为消极情绪。

总之，歌词和诗歌模型提供了强大的结果，后者可能更擅长负面情绪检测。6-Emotions 模型显示出不同的精确度和回忆，表明歌词中的情感分类方法细致入微。令人惊讶的是，最好的歌词模型超过了音频模型的结果，证实了歌词与价识别任务的相关性。

6

两种模式的融合 研究了三种不同的音频和文本结果融合方法，以利用这两种模式来增强音乐情感分析系统的性能。这些方法包括：基于最高概率的类选择、平均预测和加权组合。结果如图 ?? 所示。

第一种方法利用音频和文本模型对负类和正类的预测，选择概率最高的类。这种方法发现，在大约 19 个% 案例中，由于概率较高，选择了音频分类，而在大约 81 个% 案例中，文本分类是首选。

第二种方法涉及对每个类的两个模型的预测进行平均，并选择平均概率较高的类。有趣的是，这种平均方法的结果反映了最高概率方法的结果，表明这两种方法之间的性能没有显着差异。

第三种方法探索了音频和文本预测的不同加权组合，根据最佳平均 F 分数寻求最佳混合。音频的权重为 60 %，文本的权重为 40 %，效果最佳。与前两种融合方法相比，

加权方法在几乎所有指标上都显示出改进。具体来说，它在负效价方面表现出更高的精确度和 F 分数，而在正效价方面表现出略高的召回率和 F 分数。

总之，将音频和文本纳入音乐的情感分析比单独分析这些模态会产生更好的结果。在测试的融合方法中，具有 60 % 音频和 40 % 文本比例的加权方法成为最有效的策略。虽然文本模型在负面情绪分类方面略优于音频模型和融合方法，但音频-文本组合结果在识别积极情绪方面明显表现出色。然而，我们还想指出，许多最有趣的情况发生在音频和歌词表现出截然不同的情绪时。这通常是艺术上的故意。我们的方法也有助于发现此类情况。

7 结论 & 未来工作

在本文中，我们研究了音乐中文本和音频的情感分析，开发了一种将两种模态的结果结合起来的方法。虽然个别模型实验显示出可接受的表现，但我们发现在对负面情绪进行分类方面存在弱点。然后，我们测试了将音频和文本结果组合在一起的简单方法，最有效的是 60 个% 音频和 40 个% 文本加权，强调了考虑这两种模式的重要性。对错误分类以及音频和文本分类之间的差异的分析揭示了注释调整中的潜在错误。情绪分类学差异和必要调整的显著影响可能影响了结果。

MER 的挑战和研究的局限性包括缺乏标准化的情感分类法，对音乐的主观感知，以及专注于同时分析相同音乐作品的文本和音频的研究有限。这项工作为未来在该领域研究和开发新方法奠定了基础。

尽管存在许多挑战，但音乐情感分析是一个具有巨大潜力的复杂领域。未来的研究可以通过训练它们将歌曲分类到罗素环形模型的四个象限来改进模型，并通过开发一种模型来识别通过歌词和音频在音乐中传达的情感来增强组合模态的方法。但是，用于模型训练的数据质量至关重要。可以考虑创建一个高质量、全面的数据集，因为现有的数据集通常很小，并且专注于文本或音频，对现有注释所指的模态含糊不清，并且忽略了文本和音频情感之间的潜在矛盾。一个潜在的解决方案可能是将现有的高质量数据集分别创建为文本和音频创建，正如本文在小规模上测试的那样。鉴于大型、高质量的双峰数据集的稀缺性，这一步至关重要，无法推进音乐的双峰情感分析。

除了改进现有模型和开发结合音频和文本分析的新方法外，未来研究的一个有希望的方向是设计和训练能够单独和联合处理文本和音频输入的新型多模态模型。这种方法可能会对音乐中表达的情感提供更细致入微的理解，探索不同模式在传达情感表达时如何相互补充或对比。此外，集成文本和音频的多模态模型可以克服与单独分析这些模态相关的一些局限性，例如歌词和旋律之间的情感分类差异。通过利用多模态深度学习的进步，研究人员可以开发出能够更好地捕捉音乐元素和抒情内容在表达情感时的复杂相互作用的系统，为音乐情感识别和分析开辟新的可能性。