Reviewer #2: To reduce the interaction gap between modalities and integrate sentiment cues so that improve the performance of MSA. The author propose Frame-level Nonverbal feature Enhancement Network (FNENet). In FNENet, Vector Quantization (VQ) is utilized to transform frame-level features by training the index embeddings of each frame of acoustic and visual raw features. The experiments in this paper are adequate, but it lack of innovation and the reliance on technologies used in FNENet are not very novel. The paper's main focus is to optimize the initial representation of audio and video using the VQ strategy to capture their contextual information. However, further evaluation is necessary to determine its effectiveness since it is directly optimized on the features of the original dataset. The following are my concerns about the paper:

为了减少模态之间的交互差距并整合情感线索，从而提高 MSA 的性能。作者提出了帧级非语言特征增强网络 (FNENet)。在 FNENet 中，利用矢量量化 (VQ) 通过训练每帧声学和视觉原始特征的索引嵌入来转换帧级特征。本文中的实验是足够的，但它缺乏创新，并且对 FNENet 中使用的技术的依赖不是很新颖。本文的主要重点是使用 VQ 策略优化音频和视频的初始表示以捕获它们的上下文信息。然而，由于它是直接在原始数据集的特征上进行优化的，因此需要进一步评估以确定其有效性。以下是我对这篇论文的担忧：

1.The current dataset of three-modal sentiment analysis shows poor effectiveness of acoustic and visual features. For instance, the paper 'Analyzing Modality Robustness in Multimodal Sentiment Analysis' examines the stability of audio and video features. Therefore, while the authors' motivation is commendable, the VQ strategy proposed in the paper appears inadequate in compensating for the deficiencies of the original representation in the model training stage based on the non-textual features of the CMU-MOSEI and CMU-MOSI datasets, so the effectiveness needs to be re-evaluated.

目前的三模态情绪分析数据集，声学和视觉特征的效果较差，例如论文《多模态情绪分析中的模态稳健性分析》研究的是音频和视频特征的稳定性。因此，虽然作者的动机值得称赞，但论文中提出的 VQ 策略似乎不足以弥补基于 CMU-MOSEI 和 CMU-MOSI 数据集非文本特征的模型训练阶段原始表示的不足，因此其效果需要重新评估。

2.The model's innovation is insufficient. For instance, the authors have not made any significant improvements to the cross-attention module, which was proposed by Tsai et al. in 2019.

模型创新性不足，例如作者并未对 Tsai 等人在 2019 年提出的交叉注意力模块做出明显改进。

Answer:

非常感谢您提出的专业意见，很抱歉我们没有在论文中明确表明所使用的交叉注意力模块与MulT （Tsai et al. 2019）的重要差异。我们已经在论文中做出调整，详见论文方法标红的部分。

论文中：

我们的动机类似于MulT （Tsai et al. 2019），交叉注意力模块用于探索非语言模态与文本模态交互，。但与MulT中的交叉注意力模块不同的是，我们的FNENet并未在交叉注意力模块中应用未来掩码来将非语言模态语义翻译到文本模态。这种未来掩码与机器翻译领域中解码器里的下三角掩码矩阵相似。然而，我们利用嵌入TCN的CTC机制控制从非文本模态到文本模态的单向的信息交互，而并非使用未来掩码。TCN网络更能捕捉长距离的上下文关系，CTC机制。CTC-T将不同时间跨度的非语言模态和文本模态进行单向语义翻译。

3.For CMU-MOSI, CMU-MOSEI, and CH-SIMS, the authors discussed feature extraction in experimental part. The description of this part is the same as original papers of the three datasets. Therefore, the word 'elaborate' is not appropriate.

对于 CMU-MOSI、CMU-MOSEI 和 CH-SIMS，作者在实验部分讨论了特征提取。这部分的描述与这三个数据集的原始论文相同。因此，“详尽”一词并不合适。

Answer:

非常感谢您的敏锐的洞察力和提出的专业意见，很抱歉我们的用词十分不准确，给您带来了不便。我们已经在论文中做出调整，详见论文方法标红的部分。同时，我们还对全文内容的用词进行仔细的检查，以确保论文的专业性。

4.The experimental results from the t-SNE visualization experiment provide no evidence that the VQ module can capture the contextual information of sentiment, and from the perspective of feature distribution, the three modalities show the same trend after VQ, which also leads to the loss of the characteristic information of each modality. The specific information is very important in sentiment analysis, particularly in dealing with sarcasm and other semantic nuances.

t-SNE可视化实验的实验结果并没有证明VQ模块能够捕捉到情绪的语境信息，而且从特征分布来看，经过VQ后，三个模态呈现出相同的趋势，这也导致了各个模态的特征信息的丢失。而这些具体信息在情绪分析中非常重要，特别是在处理讽刺等语义细微差别时。

5.In Section 2.1, my question is why it is emphasized that "However, we do not employ the additional loss to enhance the model's learning on aligned data", and if there are similarities and differences between the authors' proposed model and the BBFN, please explain them clearly in the paper.

我的问题是，在第 2.1 节中，为什么强调“但是，我们并没有使用额外的损失来增强模型在对齐数据上的学习”，如果作者提出的模型与 BBFN 之间存在异同，请在论文中清楚地解释。

6.In Section 4.2, the authors compare the performance of the three publicly available datasets with the baseline, and the results show an improvement in the performance of FNENet. However, regarding the conclusion "our FNENet eliminates differences between modalities, promotes the fusion between modalities, and fully uses nonverbal modality information ", this may need to be demonstrated by ablation experiments.

在 4.2 节中，作者将三个公开可用的数据集与基线的性能进行了比较，结果显示 FNENet 的性能有所提高。但是，关于“我们的 FNENet 消除了模态之间的差异，促进了模态之间的融合，并充分利用了非语言模态信息”的结论，这可能需要通过消融实验来证明。