Computer Science and Application 计算机科学与应用, 2023, 13(4), 923-931
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134090

Hans汉斯

基于改进Self-MM模型的多模态情感分析

马健兵，沈琪瀚*，崔翔浩

成都信息工程大学计算机学院，四川 成都

收稿日期：2023年3月22日；录用日期：2023年4月21日；发布日期：2023年4月28日

摘　要

早期情感分析依托于神经网络在文本、图像或者音频等单个模态做情感分析，虽然在各自模态已经有了不错的效果，但是仅仅通过单模态做情感分析无法充分表达人们的情感，所以本文结合多个模态的信息应用于情感分析领域。该领域中Self-MM模型已经有了较好的实验效果，但是该模型在优化器层面还有提升的空间，本文在此基础上继续做研究，采用更先进的AdamW优化器，在公开数据集CMU-MOSI进行验证，实验结果在Acc-7、Acc-2两个分类精度上分别有0.12%和0.43%的提升。

关键词

多模态，情感分析，神经网络

Multimodal Sentiment Analysis Based on Improved Self-MM Model

Jianbing Ma, Qihan Shen*, Xianghao Cui

School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan

Received: Mar. 22nd, 2023; accepted: Apr. 21st, 2023; published: Apr. 28th, 2023

Abstract

Early sentiment analysis relies on neural networks to do sentiment analysis in individual modalities such as text, image or audio, and although there have been good results in each modality, it is not possible to fully express people's emotions by only doing sentiment analysis in a single modality, so this paper combines information from multiple modalities to apply to the field of sentiment analysis. The Self-MM model in this field has had good experimental results, but the model has room for improvement at the optimizer level. This paper continues to do research on this basis using the more

*通讯作者。

文章引用: 马健兵, 沈琪瀚, 崔翔浩. 基于改进 Self-MM 模型的多模态情感分析[J]. 计算机科学与应用, 2023, 13(4): 923-931. DOI: 10.12677/csa.2023.134090

advanced AdamW optimizer, and validates it in the public data set CMU-MOSI, and the experimental results have an improvement of 0.12% and 0.43% in the classification accuracy of Acc-7 and Acc-2, respectively.

Keywords

Multimodal, Sentiment Analysis, Neural Networks

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

情感分析旨在通过信息化、系统化的方式量化评估情感状态的工作，是人工智能的重要课题之一[1]。受限于计算能力和人工智能，处于起步阶段，早期的研究方法主要集中在通过文本做情感分析，可以分析博客、微博等社交平台上用户发表的言论，进而跟进舆情。然而，在互联网发展以及信息化基础建设的背景下，人们不再受限于通过文本的形式表达情感，还可以采用图像、音频、视频等多媒体媒介作为载体去表达自己的情绪和思想，与此同时，仅仅通过文本这一个模态难以表达出复杂的情感信息[2]。于是，采用多模态数据做情感分析的方式孕育而生。

多模态情感分析能够结合不同维度的模态信息实现信息互补，使得情感的表达更加多元化，更加贴切人们真实的情感，与此同时，由于异构的模态输入来源，如何去有效利用整合这些多模态数据源是一项重大的挑战。文献[3]率先尝试通过三模态去解决情感分析的问题，但是只是简单通过模态拼接，再使用隐马尔科夫链去完成分类任务。文献[4]通过训练单模态和多模态任务，设计了一种自监督学习策略(Self-MM)，通过单独学习的特征和设计的加权策略来平衡不同子任务的学习进度，并通过自监督学习策略生成单模态标签来减少人工标注，在多模态情感分析任务上取得了显著的效果，但是该方案在优化器层面还有可以提升的空间。本文基于 Self-MM 模型，采用了更加先进的优化器 AdamW，通过实验验证该方法效果更好。

综上所述，本文主要贡献有以下两点：

1) 采用了一种改进优化器的方案用作用于 Self-MM 模型，提高了模型的精度。

2) 梳理整理了情感分析领域，从单模态情感分析到多模态情感分析的演化进程。

2. 相关工作

2.1. 单模态情感分析

基于文本模态的情感分析，相较于图像和音频模态的数据源，文本模态存在更高语义维度的信息，也是情感分析的重要领域。从早期通过情感词典的方式，文献[5]提出了 WordNet-Affect 方法，文献[6]贡献了 BosonBLP 方案，这些都是通过建立词语与情感标签的映射关系，进而实现情感匹配。随后，基于机器学习的方案提出来，例如支持向量机 SVM [6]、贝叶斯等方法，给每个数据集贴上标签，依赖于监督学习，再去训练情感分类器识别文本中的情感信息[7]。近年来，研究方向聚焦于通过卷积神经网络[8]、LSTM [9]、注意力机制等深度学习方案去做情感分析，能够更加精细化准确预测出情感状态。

基于音频模态的情感分析，主要涉及到三个特征的识别，分别是频谱特征、音频特征和音律特征。

文献[8]采用 LSTM 和 RNN 模型融合音频信息中的韵律特征和语义特征，取得了良好的效果。文献[10]结合了 CNN 提取频谱图特征以及将 MFCC 等传统信息输入到 LSTM 中做特征提取，最后采用注意力机制做特征融合和情感分类，得到了显著的效果。

基于图像模态的情感分析，通过面部表情是最为直观的情感表达方式，具有丰富的感染力。文献[11]提出的面部动作编码系统为后续的图像情感识别研究奠定了基础，采用的 CNN 做特征提取，取得了较好的准确率。文献[12]提出了一种在 Inception ResNet 提取了特征以后，再采用 3D CNN 融合 LSTM 网络的识别方案做情感分类，得到了显著的提升。

综上所述，这三个单模态情感分析的方式在各自领域都有较好的发展，但是同时也是受限于单个模态信息的原因，单个模态的信息是无法满足多模态数据源场景的，所以本文侧重于多模态情感分析的研究。

2.2. 多模态融合技术

按照融合的时机做划分，可以划分为三种类型：早期融合、晚期融合以及混合融合。早期融合通常采用向量相加、向量拼接后者向量相乘的方式操作。首先把各自模态的特征提取出来后，再做一个串联的操作，最后把总的特征向量输出到分类器中，流程如图 1 所示。

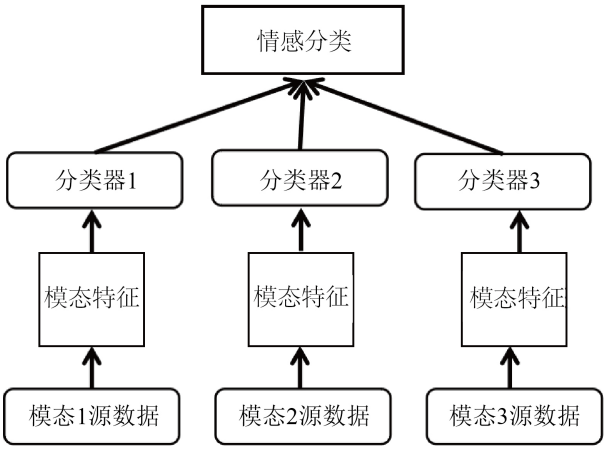

Figure 1. Early integration methods
图 1. 早期融合方式

该种方案避免了特征的归一化和缩放，能够利用好模态间的互补信息，但是存在忽略掉模态间的动态联系。随后，还有晚期融合的方式，在各个模态做完分类任务之后，再由决策层通过投票等方式得到最终的决策结果，流程如图 2 所示。

还有一种结合了早期融合和晚期融合的方式，继承了两种方式的优点并且还优化了彼此的不足，如图 3 所示，模态 1 和模态 2 两个模态率先做信息融合，结果再由分类器 1 输出，之后再于模态 3 经过分类器 2 输出的结果再做一次决策融合，进而得到最后的情感结果，这种融合方式虽然结合了模态间数据的差异性，也考虑模态间的交互性，但是也给模型增大了复杂度和可实现性。

从融合方式角度出发，可以分为两种方式。融合方式是指将各个模态的向量信息通过首尾拼接或者加权求和的方式完成向量的拼接融合工作。得益于神经网络具备良好的非线性映射能力，以及配置神经元的层级结构等方式，在特征融合方面有比较好的效果。本文采用的模型基于第三种混合融合范式去做后续的情感分析。

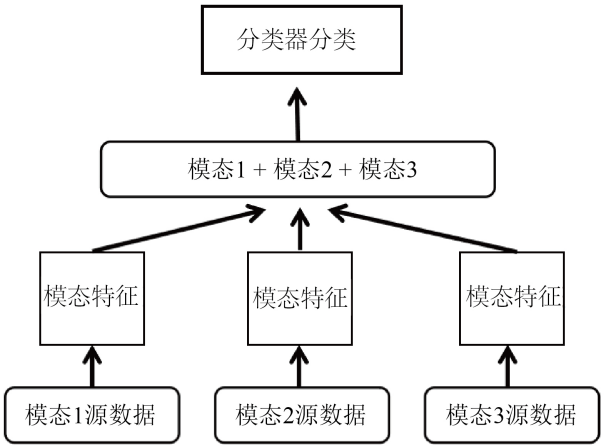

Figure 2. Decision-making period integration methods
图 2. 决策期融合方式

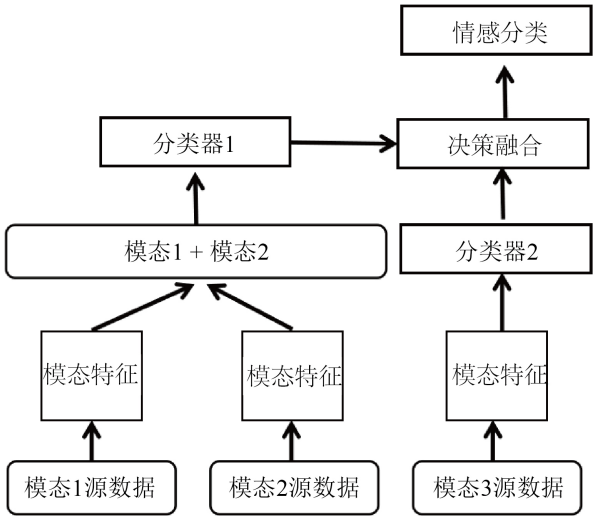

Figure 3. Hybrid integration methods
图 3. 混合融合方式

2.3. 多模态情感分析

通过多模态数据，将各个模态信息做互补会带来更好的情感分析结果。文献[13]提出了一项多模态情感分析调研，用于情感分析、情绪分析。文献[14]提出了一种基于多头注意力的多模态情感分类模型，但是存在没有融合上下文信息的问题。文献[15]提出了一种数据集合决策级的融合方法，通过张量将文本、音频和视频嵌入到决策层。上诉模型方法的侧重点在于模型的融合方式，而缺乏对于优化器层面的考虑，本文从优化器层面作为切入点做接下来的研究。

2.4. 模型优化器

Lion [16]优化器，相比较主流的优化器，具备更快的速度以及更小的内存。Adan [17]优化器，由 Sea AI LAB 和 ZERO Lab 团队共同提出，可以应用于 CV、NLP 等多个场景中。Adam [18]优化器，由 SGDM [19]和 RMSProp [20]所构成，擅长于解决一些随机小样本、自适应学习率这些问题。AdamW [21]优化器，是 Adam 优化器的变体，擅长处理大数据集，可以自动调整学习率自动调整权重衰退的系数，进而增强

模型的稳定性，避免模型过拟合。本文将采用以上优化器做后续的对比实验。

3. 模型方法

本文研究课题基于 Self-MM 模型[4]，并在其基础上，在优化器层上做优化处理，并对照目前主流的多模态情感识别模型做实验对比以及验证。Self-MM 模型结构如图 4 所示。

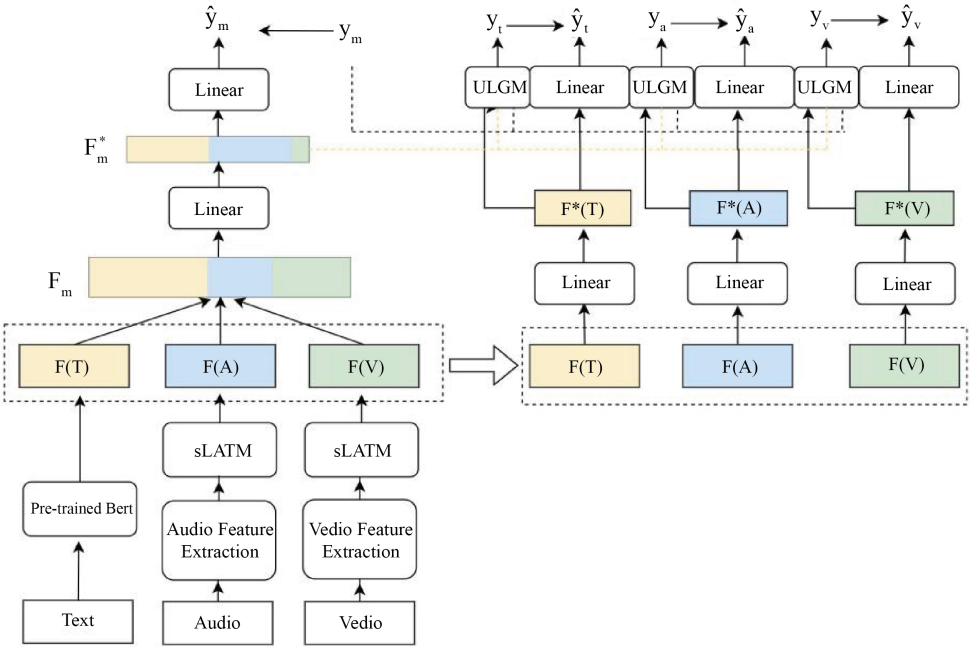

Figure 4. Self-MM model diagram
图 4. Self-MM 模型图

3.1. 特征提取层

在多模态融合之前，需要将各个模态的特征做提取。设定文本模态输入为 X_t，音频模态输入为 X_a，视频模态输入为 X_v，输入集合记为 $X=\{X_t,X_a,X_v\}$。经过特征提取后，需要得到文本特征记为 F_t，音频特征记为 F_a，视频特征记为 F_v，特征集合记为 $F=\{F_t,F_a,F_v\}$。

文本模态，本文采用 BERT [22]作为特征提取器。模型架构如图 5 所示。

音频以及视频模态，因为数据具有时序性，所以采用的 LSTM [23]作为特征提取器。LSTM 包含了三个重要的模块，分别是输入门、遗忘门以及输出门，输入门负责更新细胞的状态，遗忘门负责管理输入到改模块的信息是保留还是丢弃，输出门负责更新细胞传送给下一个神经元。

在特征提取后，这些特征向量将会应用于后续的 ULGM [4]以及多模态融合模块。

3.2. 多模态自监督和优化器模块

由于三个模态的数据维度存在差异。需要将通过向量投射的方式将三个模态的数据映射到同一个空间。映射后的特征向量记作 F_m^*，转化过程公式(1)所示。

$$F_m^*=\mathrm{ReLU}\left(W_mF_m+b_m\right) \tag{1}$$

之后，需要用到两个部分的数据信息，第一部分是由三个模态拼接而成特征 F，第二部分需要采用

ULGM 模块(Unimodal Lable Generation Module)对单个模态做自监督训练，进而得到各自模态的自监督伪标签，用于辅助后续的模型训练。

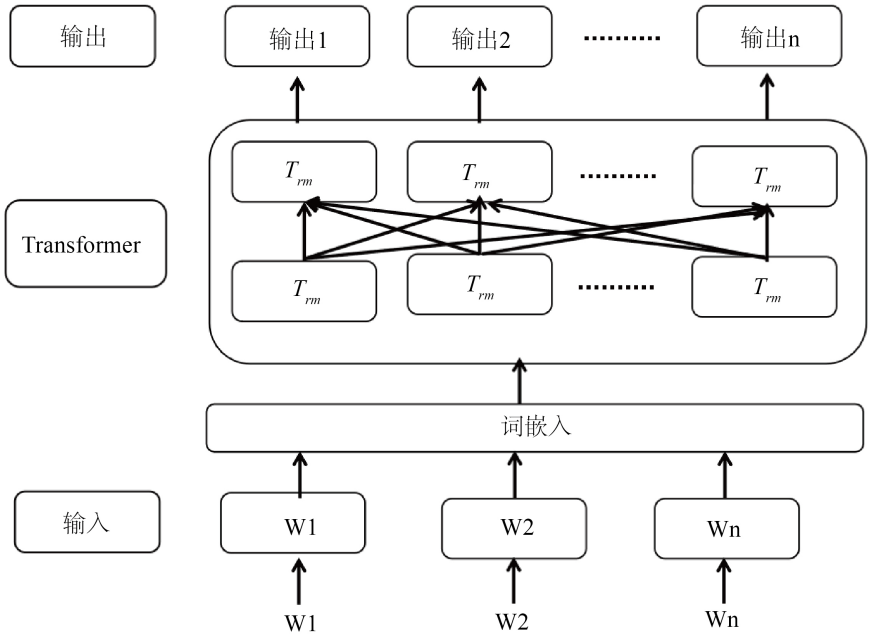

Figure 5. Self-MM model diagram
图 5. BERT 模型图

最后，我们在优化器模块做调整。优化器是用于优化损失函数，使得损失函数能够朝着更小化的方式调整训练的参数，通过实验对比本文最终采用 AdamW 作为模型的优化器，并与 Lion、Adan、Adam 优化器作对比实验。

4. 实验结果及分析

4.1. 实验环境

本文实验环境如下所示：所有实验在 Dell Precision 7920 塔式工作站上进行的，其硬件配置如下。内存空间大小 125 GB，CPU 搭载 Intel(R) Xeon(R) Silver 4210 芯片，GPU 型号为 NVIDIA TITAN RTX，显存大小 24 GB。软件环境如下，操作系统采用 Ubuntu 16.04 LTS 64 位，开发语言采用 Python 3.9 版本，深度学习框架采用 PyTorch 1.1.2 版本。

4.2. 数据集和评价指标

4.2.1. 数据集

本文采用卡内基梅隆大学公开的 CMU-MOSI 数据集[24]。这些视频的片段采集至 YouTube 上的独白集合，这个数据集由 2199 个包含意见的视频片段组成，其中电影的内容包含陈述者对于电影主题的建议和想法。CMU-MOSI 共有 93 个视频，共计 89 个发言人，其中每个片段都是在[−3, 3]的范围内以负面/正面的情绪标签标记，共计 7 个类别，分别是 Strongly Positive (标记为+3)、Positive (标记为+2)、Weakly Positive (标记为+1)、Neural (标记为 0)、Weakly Negative (标记为−1)、Negative (标记为−2)、Strongly Negative (标记为−3)。数据集按照 7 分类标签划分后如图 6 所示。

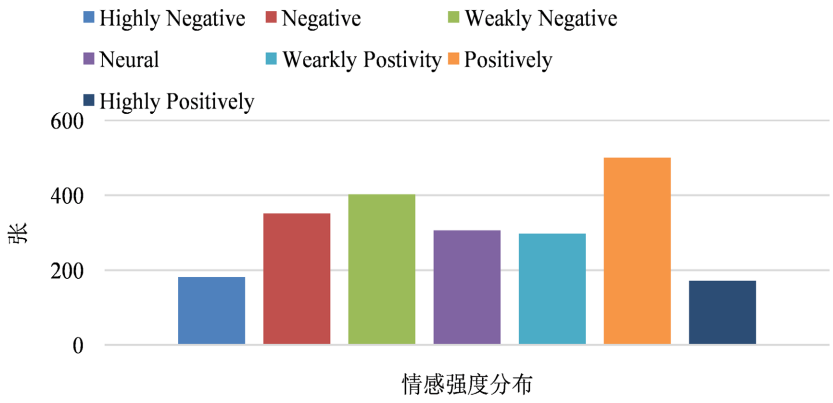

Figure 6. Data set distribution map
图 6. 数据集分布图

4.2.2. 评价指标

根据 CMU-MOSI 数据集的标签由消极到积极的程度是按照线性划分。情感分析任务可以看作是一个分类任务。本文主要采用的指标有七分类(Acc-7)、二分类(Acc-2)、平均绝对误差(MAE，公式如公式(2)所示)、相关系数(Corr，公式如公式(3)所示)。

$$\mathrm{MAE}=\frac{\sum_{i}^{n}\left|y_{i}-y_{i}^{p}\right|}{n} \tag{2}$$

$$\mathrm{Corr}=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\left(x-\bar{x}\right)^{2}\sum\left(y-\bar{y}\right)^{2}}} \tag{3}$$

4.3. 实验结果

4.3.1. 对比模型

为了验证我们改进后的模型具备提升的效果，我们选取了：

1) EF-LSTM：经典的早期融合方案，首先把特征层的向量信息做拼接操作，最后再使用 LSTM 模型做情感分类任务。

2) LF-LSTM：类似 EF-LSTM 模型，但是融合的时机不同，采用晚期融合的策略，将多个 LSTM 网络学习到的不同模态的信息做拼接，最后再分类输出预测。

3) GRAPH_MFN：通过一个简单的图神经网络在非对齐序列中学习各个模态的信息。

4) MUILT：通过 Transformer 实现跨模态之间的学习，进而更加有效地实现多模态融合任务。

5) MISA：采用编码器模块将输入模态分为不变特征和特定特征，最后融合这两种特征进行情感分析

6) Self-MM：通过自监督的方式，给每一个模态生成对应的模态标签。

4.3.2. 实验结果与分析

我们将 Self-MM 作为基础模型，在此基础上，采用不同的优化器做对比实验。分别采用 Lion、Adan、Adam 以及本文采用的 AdamW。

由表 1 我们可以看出，本文采用的 AdamW 优化器在 Acc-7、Acc-2 和 Corr 三个指标上都有所提升，并且 MAE 系数更小，说明改方案使得模型更加稳定。

我们采用 4.3.1 章节的 6 种多模态情感分析模型做对比实验，实验结果如表 1 所示。

Table 1. Comparison experiments of different optimizers
表 1. 不同优化器的对比实验

模型	Acc-7	Acc-2	MAE	Corr
Lion	40.85	80.06	0.829	0.735
Adan	43.93	84.18	0.739	0.782
Adam	45.63	84.35	0.720	0.785
AdamW	**45.75**	**84.78**	**0.721**	**0.791**

Table 2. Model comparison results on the CMU-MOSI dataset
表 2. CMU-MOSI 数据集上的模型对比结果

模型	Acc-7	Acc-2	MAE	Corr
EF-LSTM	35.39	78.48	0.948	0.669
LF-LSTM	34.52	78.63	0.954	0.658
GRAPH_MFN	34.64	78.35	0.955	0.648
MULT	36.91	80.98	0.879	0.702
MISA	41.37	83.54	0.776	0.778
Self-MM	45.63	84.35	0.720	0.785
Ours	45.75	84.78	0.721	0.791

由上表 2 可以看到，我们改进后的模型比 Self-MM 模型在 Acc-7、Acc-2 两个分类精度上分别有 0.12%和 0.43%的提升，并且在相关系数上有所提升，并且与其他主流模型作对比都有更好的实验表现，验证了我们改进方案的有效性。得益于我们采用 AdamW 作为优化器，相比较 Self-MM 采用的 Adam 等优化器，对学习率更高敏感，实现了 weight decay 解耦，效果更好。

5. 结束语

多模态情感分析与单模态情感分析相比较，多模态情感分析可以在多个感知模态(如图像、语音、文本等)中获取信息，从不同角度全面了解人的情感状态，比单一模态情感分析更加准确和全面。本文采用更加高效的优化器 AdamW 作用于 Self-MM 模型，使得模型更加稳定，避免模型过拟合，效果更好。

在接下来的工作当中，我们将会把工作重点聚焦于跨模态间的学习任务，进而提高多模态数据间的利用率，同时考虑多模态情感分析模型在鲁棒性方面的表现。

基金项目

本研究得到了四川省重点研发计划的支持(NO. 2023YFS0192 to Jianbing Ma)。

参考文献

[1] 张亚洲, 戎璐, 宋大为, 等. 多模态情感分析研究综述[J]. 模式识别与人工智能, 2020, 33(5): 426-438.

[2] Soleymani, M., Garcia, D., Jou, B., *et al.* (2017) A Survey of Multimodal Sentiment Analysis. *Image and Vision Computing*, **65**, 3-14. https://doi.org/10.1016/j.imavis.2017.08.003

[3] Morency, L.P., Mihalcea, R. and Doshi, P. (2011) Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. *Proceedings of the* 13*th International Conference on Multimodal Interfaces*, Alicante, 14-18 November 2011, 169-176. https://doi.org/10.1145/2070481.2070509

[4] Yu, W., Xu, H., Yuan, Z., *et al*. (2021) Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 10790-10797. https://doi.org/10.1609/aaai.v35i12.17289

[5] Strapparava, C. and Valitutti, A. (2004) WordNet-Affect: An Affective Extension of WordNet. *International Conference on Language Resources and Evaluation*, Vol. 4, 1083-1086.

[6] Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), **2**, 1-27. https://doi.org/10.1145/1961189.1961199

[7] Pang, B. and Lee, L. (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, **2**, 1-135. https://doi.org/10.1561/1500000011

[8] LeCun, Y., Bottou, L., Bengio, Y., *et al*. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[9] Shi, X., Chen, Z., Wang, H., *et al*. (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Vol. 1, 802-810.

[10] Luo, Z., Xu, H. and Chen, F. (2019) Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network. *AffCon@ AAAI*, 80-87. https://doi.org/10.29007/7mhj

[11] Breuer, R. and Kimmel, R. (2017) A Deep Learning Perspective on the Origin of Facial Expressions. ArXiv: 1705.01842.

[12] Hasani, B. and Mahoor, M.H. (2017) Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, 21-26 July 2017, 30-40. https://doi.org/10.1109/CVPRW.2017.282

[13] Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion*, **37**, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

[14] Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E. and Morency, L.P. (2018) Multi-Attention Recurrent Network for Human Communication Comprehension. 32*nd AAAI Conference on Artificial Intelligence* (*AAAI*-2018), New Orleans, 2-7 February 2018, 5642-5649. https://doi.org/10.1609/aaai.v32i1.12024

[15] Sun, J., Yin, H., Tian, Y., *et al*. (2021) Two-Level Multimodal Fusion for Sentiment Analysis in Public Security. *Security and Communication Networks*, **2021**, Article ID: 6662337. https://doi.org/10.1155/2021/6662337

[16] Chen, X., Liang, C., Huang, D., *et al*. (2023) Symbolic Discovery of Optimization Algorithms. ArXiv: 2302.06675.

[17] Xie, X., Zhou, P., Li, H., *et al*. (2022) Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models. ArXiv: 2208.06677.

[18] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. ArXiv: 1412.6980.

[19] Leitão, P.J., Schwieder, M. and Senf, C. (2017) sgdm: An R Package for Performing Sparse Generalized Dissimilarity Modelling with Tools for Gdm. *ISPRS International Journal of Geo-Information*, **6**, Article 23. https://doi.org/10.3390/ijgi6010023

[20] Dauphin, Y., De Vries, H. and Bengio, Y. (2015) Equilibrated Adaptive Learning Rates for Non-Convex Optimization. *Advances in Neural Information Processing Systems*, **28**, 1504-1512.

[21] Loshchilov, I. and Hutter, F. (2017) Decoupled Weight Decay Regularization. ArXiv: 1711.05101.

[22] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ArXiv: 1810.04805.

[23] Staudemeyer, R.C. and Morris, E.R. (2019) Understanding LSTM—A Tutorial into Long Short-Term Memory Recurrent Neural Networks. ArXiv: 1909.09586.

[24] Zadeh, A., Chen, M., Poria, S., *et al*. (2017) Tensor Fusion Network for Multimodal Sentiment Analysis. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, September 2017, 1103-1114. https://doi.org/10.18653/v1/D17-1115