Operations Research and Fuzziology 运筹与模糊学, 2023, 13(6), 7346-7358
Published Online December 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.136722

Hans汉斯

基于ViLT的社交媒体领域图文情感分析方法

杨　靖

上海工程技术大学电子电气工程学院，上海

收稿日期：2023年10月24日；录用日期：2023年12月20日；发布日期：2023年12月28日

摘　要

现有的图文情感分析方法更多地集中于图文信息的特征提取方面，较少关注不同模态之间的特征对齐，针对这一问题提出了一种基于ViLT (Vision-and-Language Transformer)的社交媒体领域图文情感分析方法。结合社交媒体文本长度较短、语法不规范等特点，选用BERTweet作为文本编码器，利用ViLT模型将图片切片投影的方法提取图像特征。将文本特征与图像特征进行拼接，送入同一个Transformer模块，得到基于图文多模态分析的情感结果。并充分挖掘文本与图像自身的特征得出两个基于单模态的情感分析结果，最后对三种情感分析结果使用加权融合策略确定最终的情感极性。该方法在公开数据集上进行了实验，验证了本文情感分类方法的有效性。

关键词

跨模态，情感分析，注意力机制，特征融合

Image-Text Sentiment Analysis Method in Social Media Domain Based on ViLT

Jing Yang

School of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai

Received: Oct. 24th, 2023; accepted: Dec. 20th, 2023; published: Dec. 28th, 2023

Abstract

The existing image-text sentiment analysis methods focus more on feature extraction of image and text information with less attention to feature alignment between different modalities. Therefore, this paper proposes an image-text sentiment analysis method in the social media domain based on Vision-and-Language Transformer (ViLT). Combining the features of short length and irregular syntax of social media texts, BERTweet is chosen as the text encoder and image features are extracted by slicing and projecting images using ViLT model. The text features and image features

文章引用: 杨靖. 基于 ViLT 的社交媒体领域图文情感分析方法[J]. 运筹与模糊学, 2023, 13(6): 7346-7358.
DOI: 10.12677/orf.2023.136722

are stitched together and sent to the same Transformer module to get the sentiment results based on the multimodal analysis of the graphical text. And the features of text and image themselves are fully exploited to derive two unimodal-based sentiment analysis results. Finally, the final sentiment polarity is determined using a weighted fusion strategy for the three sentiment analysis results. The method is experimented on a public dataset to verify the effectiveness of the sentiment classification method in this dissertation.

Keywords

Cross-Modal, Sentiment Analysis, Attention Mechanism, Feature Fusion

1. 引言

情感传递方式是多样的，人们可以使用多种形式共同表达情感态度，以提供更全面的情感信息。随着移动终端技术的不断改进和 5G 通信的发展，人们表达情感的方式也更加多元化。用户在种类繁多的社交媒体平台发布内容，其形式包括文字、图像、语音，同时还借助视频、表情符号、颜文字等，以多种模态组合的方式表达观点，丰富情感层次。多模态情感分析研究应运而生。多模态情感分析指的是利用两个或多个信息来源，例如文本、语音、图像等，对情感进行分析的方法。与单一文本传递信息的方式相比，利用多种媒体形式能够更好地表达情感，让情感传递更加准确和有力。

据 2015 年针对推特用户的一项统计调查[1]，超过三分之二的被调查者表示会在发布推文时添加图片，以增强情感表达。这进一步说明了多模态情感传递的普遍性和优势。现有的图文模态情感分析方法大多采用双流结构[2]，对文本和图片分别提取特征向量，独立地输入到两个不同的特征处理模块，充分挖掘单模态中的信息再进行特征融合，而忽略了图文的关联性。在社交媒体用户所发布的内容中，通常会选择将图片作为一种补充或增强文本所传达情感的载体，但随着社交语言的发展，反讽、图文无关等现象开始广泛出现，同一推文中不同模态所表达的情感不再具有一致性。

本文针对上述问题，设计了一种基于 ViLT 的社交媒体领域图文情感分析方法(Image-text sentiment analysis in social media domain based on ViLT, SoD-ViLT)。将文本特征与图像特征进行拼接，送入同一个 Transformer 模块，得到基于图文多模态分析的情感结果，并利用文本与图像自身的特征得出基于单模态的情感分析结果，对三种情感分析结果使用加权融合策略确定最终的情感极性。针对图文情感分析任务和图文方面级情感分析任务分别进行了实验，在公开数据集上验证了本文情感分类方法的有效性。

2. 相关工作

2.1. 多模态图文情感分析

随着社交媒体的快速发展，多模态信息量呈爆发式增长。不同来源的信息，信息的不同形式都可以单独称作一种模态。作为人类情感载体的信息模态有文本、图像、语音、视频、姿态等。不同模态的提供的信息不完全相同，其特征中体现着来自维度和层次差异的情感信息，将不同模态信息进行融合是多模态情感分析的核心。多模态融合是从各种来源接收的数据中过滤、提取和组合所需特征的过程。常见

的有早期融合，后期融合和中期融合。

早期融合也称为特征融合，将多个独立的输入模态特征组合在一起形成联合特征向量，如文本、音频或视觉的所有特征组合成单个特征向量，然后将其输入到分类算法中。Xu 等人[3]提出了一个充分考虑图像和文本之间相互作用的共记忆网络，使用两个内存网络来处理两种模态的数据，利用协同记忆注意机制，使用文本信息来查找视觉关键内容，使用图像信息来定位文本关键字，将最后查找到的关键内容与关键字特征表示连接起来进行情感分类。Zadeh 等人[4]提出一种多注意力循环网络 MARN，对特定于视图的动态以及跨视图动态连续建模。其中两个关键组成部分为长短期混合记忆 LSTHM 和多注意力块 MAB，针对每种模态，使用 LSTHM 对特定于视图的动态进行建模，在每一个时间步中，使用 MAB 识别不同的跨视图动态，并根据所有 LSTHMs 的隐藏状态信息对输出维度进行加权。

在后期融合中，首先对每种模态的特征进行独立处理和分类，然后融合分类结果形成最终的决策向量，从而预测情感，因此这个过程也被称为决策层融合。Cai 等人[5]采用了一种多个卷积神经网络联合的框架，Multi-CNN 以文本卷积神经和图像卷积神经网络联合表示作为输入，分别提取两种表示，通过捕获文本和图像的组合信息，实现了对多媒体情感分析的有效性能。Huang 等人[6]为了自动关注与情感最相关的区域和重要词汇，利用视觉注意机制自动聚焦图像中情感区域，利用语义注意机制突出文本中最具情感的词语，采用多层感知器来挖掘不同模态特征之间的非线性相关性，最后提出了视觉、语义和多模态注意模型的后期融合方案，以获得情感分类的最终决策。

中期融合又称混合融合，是特征层融合和决策层融合两种策略的结合。You 等人[7]提出了一个新的端到端联合视觉文本情感分析框架，以结构化的方式集成文本和视觉信息。在树状结构的 LSTM 中引入了注意力机制，以学习图像区域与描述性单词之间的对齐或对应关系。在树的根节点上获得联合特征，并将其提供给多层感知器用于训练情感分类器。同时还引入了一个辅助任务视觉文本语义嵌入，以帮助注意模型的学习。凌海彬等人[8]构造了两类特征，内容特征即微博中除文本以外对情感具有指示作用的隐藏信息，和体现单个用户情感表达特点的用户特征，然后将微博句子和内容特征、用户特征表示为向量矩阵输入到多特征融合的文本情感分类模型中实现文本的情感分类，结合基于迁移学习的图片情感分类模型，最后将文本和图片情感分类模型进行融合。蔡宇扬等人[9]提出模态信息交互模型，自适应特征融合去除冗余信息，根据全局特征选择包含情感信息的特征。

基于注意力机制，多模态图文情感分析向着高维度特征融合方向发展，高层特征可以很好的表示模态间的交互信息，但丢失了单模态的内在特征。不同与以上工作，本文在图文融合特征的基础上，根据不同模态的情感可靠性，调整各模态的情感贡献度，同时关注模态内与模态间的信息。

2.2. 多模态方面级情感分析

针对文本内含有多个方面词的情感分类任务又划分多模态方面级情感分析(Multimodal Aspect-Based Sentiment Analysis, MABSA)细粒度任务，又称面向目标的多模态情感分析，或者基于实体的多模态情感分析。多模态方面情感分类旨在对抽取出的每个方面词进行情感分类。Xu 等人[10]提出了多交互记忆网络，分别考虑方面词对文本和图像的影响，该模型首先获得基于方面词引导的文本特征和视觉特征，为了学习模态间的相关性，将单模态的自注意力特征与跨模态注意力特征按位相加取平均，并采用 GRU (Gated Recurrent)单元将当前单模态的特征向量与平均后的特征向量结合起来，得到新的各模态特征向量，将最后一层 GRU 输出的文本特征和视觉特征连接，作为 softmax 的输入进行情感极性的分类。Yu 等人[11]提出了 TomBERT 模型，设计了目标－图像感知模块，将目标词的特征向量与图像特征进行多头跨模态注意力，获得目标词敏感的图像特征，再与文本特征连接，堆叠额外的编码层融合图文特征。并为两个提供了文本中目标词的多模态数据集 Twitter-15 和 Twitter-17，标注了每个目标词的情感极性。Khan 等

人[12]引入了一种全新的用于多模态情感分析的结构，将图片信息转化为文本信息，利用 Transformer 生成关于图像的描述性文本，为模型提供更多的文本信息进行辅助训练，并与目标词组合构建成一个辅助句子，最后将图片和目标词所对应的文本与辅助句子送入 BERT [13]预训练模型进行情感分类。为验证本文提出模型的通用性，在多模态方面级情感分析任务上进行实验。与上述工作相比，本文模型在建立目标词与图片关联的基础上，将图片的全局特征送入最后的分类层。

3. 模型介绍

本节详细介绍提出的基于 ViLT 的社交媒体领域图文情感分析方法，SoD-ViLT 的总体框架如图 1 所示。

模型结构可大致分为三层：特征提取层，特征处理层和情感分类层。SoD-ViLT 模型在特征提取层分别对文本和图像进行预处理并提取特征。在特征处理层，SoD-ViLT 对文本特征、图像特征和图文特征分别采用了三种不同的方法进行处理，其核心都是基于 Transformer 的 Encoder 模块，最后将三部分的情感分类结果加权融合，得到最终的情感极性。

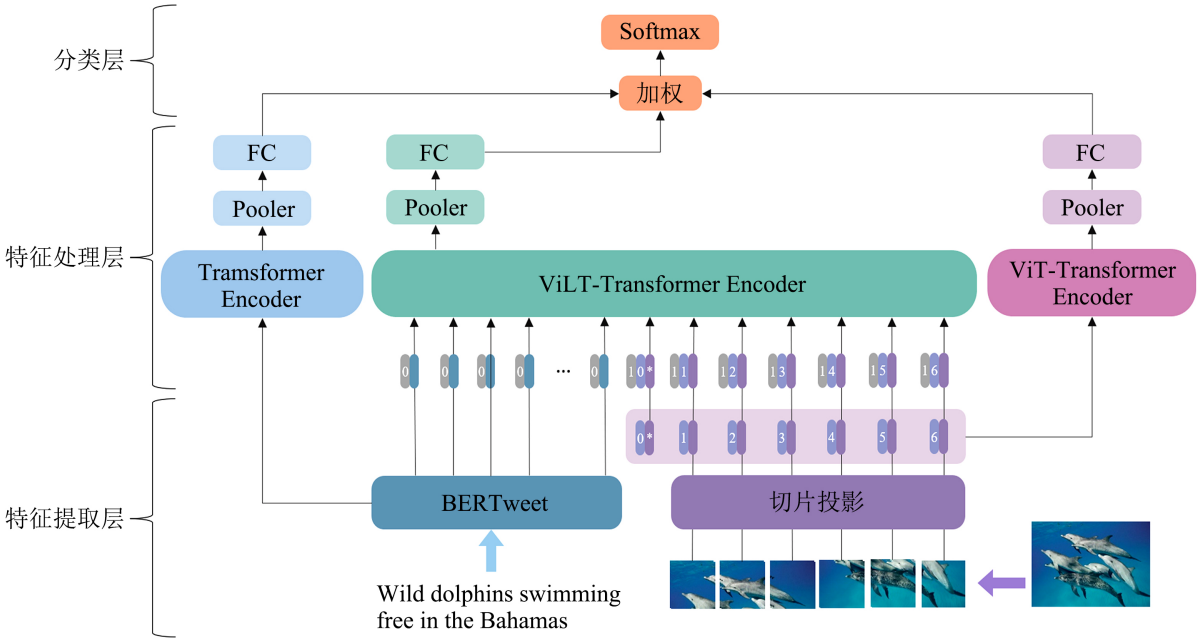

Figure 1. Architecture of SoD-ViLT
图 1. SoD-ViLT 框架图

3.1. 特征提取层

3.1.1. 文本特征提取

本文实验针对社交媒体平台的多模态数据，社交语言通常使用口语化的表达方式，且夹杂着缩写、俚语、网络用语和表情符号等元素，也包含一些语法错误或是拼写错误。同时用户的情感表达方式也更直接，情感词语的出现频率更高，且许多推文只包含一句话或一个短语。

考虑到社交文本数据有以上这些特点，不适合采用 BERT 等相似的基于大规模规范文本语料库训练出来的语言模型。本文实验采用 BERTweet [14]作为文本编码器，BERTweet 是第一个用于英语推文的大规模公共预训练语言模型，模型结构与 BERT 相同。BERTweet 的输入序列 $X=(x_0,x_1,\cdots,x_n)$ 为，其中 x_0

为句首的[*CLS*]起始标记，编码后的输出序列为$T=\left(t_0,t_1,\cdots,t_N\right)$，$N$是模型输入的最大文本长度。且编码后的序列，已添加位置嵌入信息。

3.1.2. 图像特征提取

考虑到每个样本中的图片大小不一定相同，将图片尺寸统一调整为384 × 384像素，采用与ViLT相同的图像特征提取方法，对图像切片投影，将图片切分成多个32 × 32大小的图像块，再投影成一维序列，图片特征可表示为：$I=\left(i_1,i_2,\cdots,i_{144}\right)$。增加一位用于分类的[*CLS*]标志位$i_0$，为了将图像块的排列顺序与原图片的二维信息保持一致，在图片特征I上增加图像块的位置编码I^{pos}，得到图像特征提取模块的输出特征$\overline{I}$，如公式(1)。

$$\overline{I}=\left[i_0,i_1,i_2,\cdots,i_{144}\right]+I^{pos} \tag{1}$$

3.2. 特征处理层

3.2.1. 文本特征处理

经过文本特征提取模块BERTweet预训练模型得到的特征序列，输入到标准Transformer-Encoder层，单文本特征处理结构如图2所示。

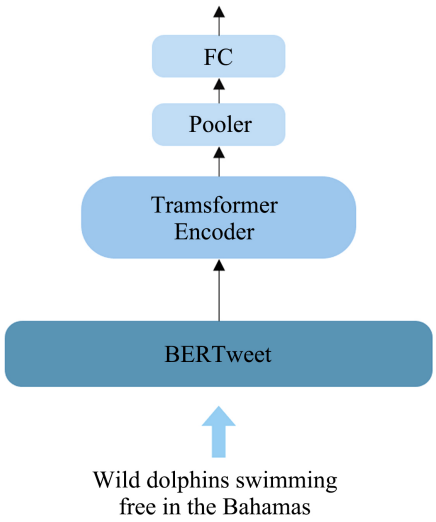

Figure 2. Structure of text feature processing
图2. 单文本特征处理结构

其中，多头注意力头数为12，隐藏层大小为768。在经过Transformer编码器内的多头注意力机制学习后，用[*CLS*]标志位的特征向量进行基于文本的情感分类，输入全连接层得到情感分类结果p_i^T。

3.2.2. 图像特征处理

图像特征$\overline{I}$在经过ViT-Transformer编码器内的多头注意力机制学习后，用[*CLS*]标志位的特征向量进行基于图像的情感分类，输入全连接层得到情感分类结果p_i^I，单图像特征处理结构如图3所示。

ViT-Transformer编码器由堆叠的块组成，其中包括一个多头自注意力层(Multi-head self-attention, MSA)和一个多层感知机(Multilayer perceptron, MLP)。ViT-Transformer的编码器与标准Transformer编码器唯一的区别是层归一化(Layer Normalization, LN)的位置，Transformer编码器的LN分别添加在MSA和MLP之后，而ViT-Transformer编码器的LN在MSA和MLP之前。

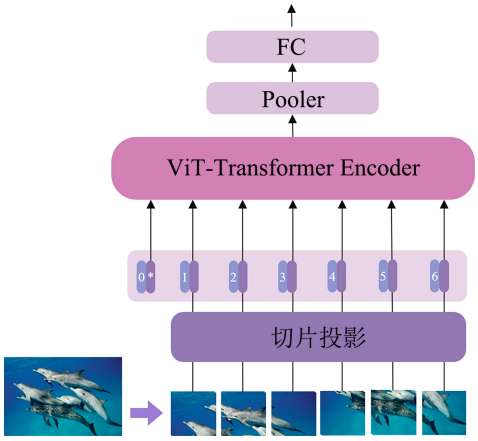

Figure 3. Structure of image feature processing
图 3. 单图像特征处理结构

图片特征的计算公式见公式(2)至公式(5)。

$$z_0 = \bar{I} = \left[i_0, i_1, i_2, \cdots, i_{144}\right] + I^{pos} \tag{2}$$

$$z_l' = MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1} \quad l = 1, \cdots, L \tag{3}$$

$$z_l = MLP\left(LN\left(z_l'\right)\right) + z_l' \quad l = 1, \cdots, L \tag{4}$$

$$p_i^I = LN\left(z_l^0\right) \tag{5}$$

其中，z_0 为编码器的输入，z_l' 表示多头注意力机制的输出，z_l 表示多层感知机的输出，z_L^0 是第 L 层输出向量的第 0 位，即[*CLS*]分类标志位，L 为编码器的层数。

3.2.3. 图文特征处理

图文特征处理使用多层 Transformer 编码器对多模态特征进行融合，得到基于图文特征的情感分类结果 p_i^{TI}，结构如图 4 所示。

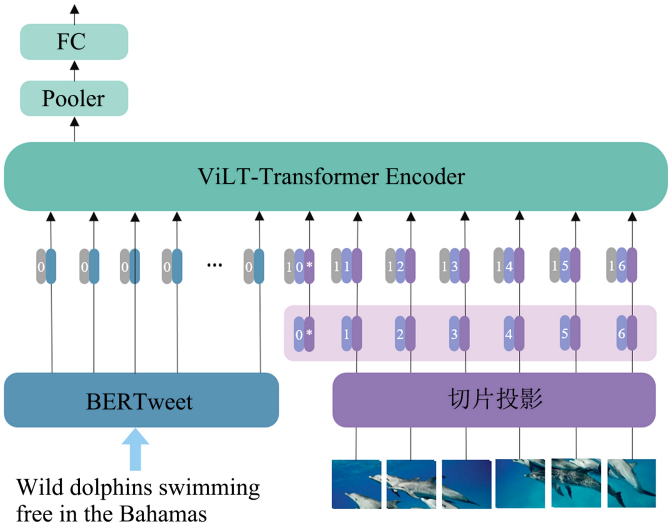

Figure 4. Structure of image and text features processing
图 4. 图文特征处理结构

为区分文本特征与图片特征，在已提取的文本特征向量 T 上添加文本类型嵌入向量 t^{type}，图片特征向量 $\overline{I}$ 上添加图片类型嵌入向量 v^{type}，如公式(6)和公式(7)所示。

$$\overline{t}=\left[t_0,t_1,\cdots,t_N\right]+t^{type} \tag{6}$$

$$\overline{v}=\overline{I}+v^{type} \tag{7}$$

将文本特征 $\overline{t}$ 和图像特征 $\overline{v}$ 进行合并，得到多模态特征 m_0，作为多模态特征输入到 ViT-Transformer 编码器中，编码公式如公式(8)至公式(11)所示。

$$m_0=concat\left(\overline{t},\overline{v}\right) \tag{8}$$

$$m_l'=MSA\left(LN\left(m_{l-1}\right)\right)+z_{l-1}\quad l=1,\cdots,L \tag{9}$$

$$m_l=MLP\left(LN\left(m_l'\right)\right)+m_l'\quad l=1,\cdots,L \tag{10}$$

$$p_i^{TI}=\tanh\left(m_L^0W_{pool}\right) \tag{11}$$

其中，m_0 为编码器的输入，m_l' 表示多头注意力机制的输出，m_l 表示多层感知机的输出，m_L^0 是第 L 层的输出向量第 0 位，即[*CLS*]分类标志位，W_{pool} 为线性投影，L 为编码器的层数。

3.3. 情感分类层

上述三种特征处理方式分别对单文本、单图像和图文模态特征进行情感分类。在情感分类层，采用加权融合的方式计算情感极性，充分利用模态的内在信息，并将不同模态的语义进行互补，结合三种模型的优势，确定最终的图文对情感类别。计算公式如公式(12)所示。

$$p_i=\left(1-\alpha-\beta\right)p_i^{TI}+\alpha p_i^I+\beta p_i^T \tag{12}$$

其中，α 为文本特征情感分类结果的权重，β 为图片特征情感分类结果的权重。

4. 实验与结果分析

4.1. 数据集

本文在 4 个数据集上进行实验，分别为图文情感分析数据集 MVSA-Single 和 MVSA-Multiple，数据集信息如表 1 所示，以及图文方面级情感分析数据集 Twitter-15 和 Twitter-17，数据集信息如表 2 所示。

MVSA 的两个数据集中每条数据是由单个图像和相应文本组成的，不同之处在于 MVSA-Single 数据集的图像和文本分别对应一个情感标签，而 MVSA-Multiple 数据集的每个图文对样本包含三个情感标注标签。MVSA-Single 和 MVSA-Multiple 的训练集、验证集、测试集划分方式，参照前人的工作[15]，按照 8:1:1 的比例随机拆分数据集。

Twitter-15 与 Twitter-17 数据集，分别由 Zhang 等人[16]和 Lu 等人[1]提出，主要包括了 2014 年至 2015 年期间和 2016 年至 2017 年期间，在 Twitter 平台上发布的多模态推文。Yu 等人[11]对每个方面词的情感进行了进一步标注，情感标签为积极、消极和中性三种。Twitter-15、Twitter-17 数据集中每个样本 D 都由上下文 S、方面词 T 和图片 I 组成。在本文实验中，方面词和上下文的输入格式参照 Yu 等人的工作，通过插入特殊标记[*CLS*]和[*SEP*]将上下文 S 和方面词 T 连接起来，文本输入格式为[*CLS*]*T*[*SEP*]*S*[*SEP*]。

Table 1. MVSA dataset statistics
表 1. MVSA 数据集统计

	积极	中性	消极	总数
MVSA-Single	2683	470	1358	4511
MVSA-Multiple	11,318	4408	1298	17,024

Table 2. Twitter dataset statistics
表 2. Twitter 数据集统计

Twitter-15					
	消极	中性	积极	总数	平均长度
训练集	368	1883	928	3179	16.7
验证集	149	679	303	1122	16.7
测试集	113	607	317	1037	17
Twitter-17					
	消极	中性	积极	总数	平均长度
训练集	416	1638	1508	3562	16.2
验证集	144	517	515	1176	16.4
测试集	168	573	493	1234	16.4

4.2. 实验环境与参数设置

本文实验基于 Pytorch 深度学习平台，Pytorch 版本为 1.13.1，Python 版本为 3.7，transformers 版本为 4.26.1。文本模态参照 BERTweet 预训练模型的参数配置初始化文本特征。对于图像模态，参考 ViLT 的图像处理参数设置，将原始图像切分成多个 32 × 32 大小的图像块，经过线性投影得到图像的特征向量。本文实验的重要参数值设置如表 3 所示。

Table 3. Parameters of the experiment
表 3. 实验参数设置

模态	参数	参数值
文本	文本最大长度	40
图像	图像块大小	32*32
多模态	学习率	2e−5
	Dropout	0.1
	批处理样本数	8
	α	0.2
	β	0.2
	Attention Heads	12

4.3. 基准方法与评估指标

本文提出了一种基于 ViLT 的图文情感分析方法，为验证该方法的有效性，针对图文情感分析任务

和图文方面级情感分析任务两个不同的任务，选择了两组基准方法。

1) 图文情感分析任务对照组

SentiBank：视觉模态分析模型，选取具有一定检测精度的 1200 个形容词名词对(Adjective-Noun Pairs, ANP)，为每个 ANP 训练一个分类器，从而实现图像分类，利用形容词－名词对来代表视觉内容的属性。自动检测图像视觉属性，为每个属性分配一个情感得分。

VggNet-19：视觉模态分析模型，采用同样大小的 3 × 3 尺寸的卷积核，2 × 2 尺寸的最大池化，共包含了 16 个卷积层和 3 个全连接层的卷积神经网络。

SentiStrenth [17]：文本模态分析模型，利用单词和短语进行情感分析，识别出不同情感的组合，并根据其权重计算出文本的总体情感极性。

LSTM-Attention：文本模态分析模型，利用 LSTM 对文本序列进行编码，然后通过 Attention 机制获取文本的上下文信息，最终将编码后的文本向量输入到全连接层中进行情感分类。

BiLSTM-Attention：文本模态分析模型，在 LSTM-Attention 基础上，利用双向 LSTM 从正向和反向两个方向对文本序列进行编码，捕捉更全面的上下文信息。

SentiBank + SentiStrenth [18]：多模态分析模型，将 SentiBank 和 SentiStrenth 的结果进行决策级的融合。

MLSA [19]：多模态分析模型，利用文本引导的多层次空间注意力对图像进行特征提取。

ViLT：多模态分析模型，将图片切分成图像块投影成一维序列，与文本特征拼接输入 Transformer 编码器融合多模态特征。

2) 图文方面级情感分析任务对照组

Res-Target：视觉模态分析方法，使用 Resnet-152 网络提取图像特征，将其与方面词的特征向量拼接，输入 BERT 模型中进行分类。

IAN [20]：文本模态分析方法，交互式注意网络模型，分别生成上下文和方面词的表示，利用注意力机制进行交互学习，最后将二者拼接预测方面词的情感极性。

RAM [21]：文本模态分析方法，多层注意力和记忆网络模型，用双向 LSTM 结构建立记忆体，通过多层注意力机制从记忆体中获得加权后的特征。

MGAN [22]：文本模态分析方法，多粒度注意力模型，将方面词整体对文本影响的粗粒度权重，以及方面词中每个单词对文本影响的细粒度权重拼接。

ViLT：图文模态分析方法，视觉和语言预训练模型，将文本特征和图像切片投影后的特征拼接输入 Transformer 编码器。

mBERT：图文模态分析方法，此模型采取了双流结构，包含图像处理模块与文本处理模块，没有针对方面词增加额外的处理模块。图像处理模块使用了 ResNet-152 模型来提取图像特征，同时采用 BERT 模型提取具有上下文感知能力的文本特征。随后，将图像特征和文本特征进行拼接，再在此基础上堆叠一层 BERT 模型，以建立视觉和文本特征之间的跨模态交互。

TomBERT：此模型以 mBERT 的文本－图像双流结构为基础，在图像处理部分设计了一个目标实体－图像交互模块，以获得带有目标实体信息的图像特征。采用与 mBERT 相同的特征融合模块，将图像特征与文本特征拼接输入 BERT 进行情感分类。

4.4. 实验结果与分析

4.4.1. 图文情感分析实验结果

表 4 展示了图文情感分析任务中不同的情感分类方法在 MVSA-Single 和 MVSA-Multiple 两个数据集

上的实验结果。实验结果表明，本文提出的基于 ViLT 的社交媒体多模态情感分析方法，在 MVSA-Single 和 MVSA-Multiple 社交媒体实验数据集上准确率和 F1 值都有一定提升，验证了本文方法的有效性。根据表 4 所展示的实验结果可以得出以下结论：

Table 4. Experimental results on MVSA datasets
表 4. MVSA 数据集实验结果

方法	MVSA-Single		MVSA-Multiple	
	准确率	F1 值	准确率	F1 值
SentiBank	45.22	43.80	50.02	51.15
VggNet-19	67.76	60.7	64.06	61.89
SentiStrenth	49.86	48.45	50.57	49.84
LSTM-Attention	65.98	64.28	67.86	66.18
BILSTM-Attention	66.8	64.79	68.97	67.03
SentiBank + SentiStrenth	52.05	50.08	65.62	55.36
MLSA	68.84	69.89	68.12	69.07
ViLT	75.9	75.3	69.9	67.3
SoD-ViLT	78.63	77.82	72.01	70.07

1) VggNet-19 模型在两个实验数据集上的准确率和 F1 值均远高于基于 SentiBank 方法的结果。这表明，在图像情感分类方面，深度神经网络相较于其他方法具有更出色的性能。在单文本数据对比试验中，基于神经网络的 LSTM-Attention 和 BILSTM-Attention 方法在准确率和 F1 值方面均优于基于语法的 SentiStrenth 方法。验证了深度神经网络在特征提取方面的强大性能。利用 SentiBank 和 SentiStrenth 相结合的图文多模态分析方法，模型性能也远低于基于神经网络的方法。

2) 单文本模态情感分析方法中的两个神经网络模型，BILSTM-Attention 和 LSTM-Attention，相较于 LSTM，双向 LSTM 结构可以更好地捕获文本语义，所以 BILSTM-Attention 模型的实验结果整体优于 LSTM-Attention 模型。

3) 相比于单模态的情感分类方法，基于多模态特征的情感分类模型在准确率和 F1 值上都有更高的实验结果，说明利用多种模态数据信息可以更准确地判断情感极性。

4) 本文提出的 SoD-ViLT 模型与 ViLT 模型相比，MVSA-Single 数据集上的准确率和 F1 值分别提升了 2.73%和 2.52%。在 MVSA-Multiple 数据集上的准确率和 F1 值分别提升了 2.11%和 2.77%。SoD-ViLT 模型在文本特征提取模块所使用的是 BERTweet 预训练模型，针对实验数据集的社交媒体领域特点，提供更匹配的初始化权重。且使用加权规则充分利用各模态信息，得到最后的情感极性。这说明 SoD-ViLT 模型在 ViLT 模型上增加或改变的模块对提高社交媒体情感分类性能是有效的。

4.4.2. 图文方面级情感分析实验结果

表 5 展示了图文方面级情感分析任务中不同的情感分类方法在 Twitter-15 和 Twitter-17 两个数据集上的实验结果。实验结果表明，本文提出的 SoD-ViLT 模型在 Twitter-15 和 Twitter-17 社交媒体实验数据集上，与基准方法相比准确率和 F1 值都有一定提升。根据实验结果可以得出以下结论：

Table 5. Experimental results on Twitter-15 and Twitter-17 datasets
表 5. Twitter-15 和 Twitter-17 数据集实验结果

方法	Twitter-15		Twitter-17	
	准确率	F1 值	准确率	F1 值
Res-Target	59.88	46.48	58.59	53.98
IAN	68.18	67.14	63.41	62.94
RAM	70.68	63.05	64.42	61.01
MGAN	71.17	64.21	64.75	61.46
ViLT	71.80	65.10	62.70	58.80
mBERT	75.79	71.07	68.80	67.06
TomBERT	76.18	71.27	70.50	68.04
SoD-ViLT	76.07	71.19	68.95	67.42

1) 单文本模态分析方法的准确率远高于基于视觉模态的 Res-Target 方法，这说明仅基于图片和方面词信息并不能有效的分析情感极性，图片只能提供辅助信息，文本是预测方面词情感极性的主要信息来源。

2) 基于多模态的方法要优于基于单模态的方法，说明图片和文本能够提供互补信息，图片对文本起到一定的辅助作用。

3) 多模态方法 ViLT 作为本文实验方法的原型基础，其各项评价指标与其他多模态基准方法相比较差，而本文提出的 SoD-ViLT 模型性能超过所有基准方法，说明了 SoD-ViLT 在 ViLT 的基础上，针对社交媒体的推文特点选择 BERTweet 作为文本编码器等改进措施的有效性。

4) mBERT 与本文提出的 SoD-ViLT 都没有关注方面词的重要性，在同样丢失方面词信息的情况下，SoD-ViLT 的性能超过了 mBERT，原因在对图片信息的处理，本文实验采用了与视觉预训练模型 ViT 相似的图像特征提取方式，生成了更具有泛化能力的图像特征。

5) TomBERT 在两个实验数据集上都取得了比本文提出的 SoD-ViLT 模型更好的结果，与 mBERT 相比 TomBERT 增加了方面词与图片匹配模块，说明建立方面词与图片之间关系的重要性，突出图片中与方面词相关区域能有效提高方面级情感分析的结果。

4.4.3. 情感分类权重比例分析

为达到模型的最优性能，本小节在 MVSA-Single 数据集上，通过网格搜索法确定文本特征情感分类结果权重系数 α，和图片特征情感分类结果的权重系数 β 的取值。由于 $\alpha\in(0,1)$，$\beta\in(0,1)$，同时 $(1-\alpha-\beta)\in(0,1)$，将步长设定为 0.1，则 $\alpha\in[0.1,0.4]$，$\beta\in[0.1,0.4]$。实验结果如图 5 和图 6 所示。

从图 5 可以看出，当 β 的值即图片特征情感分类结果的权重系数为 0.4 时，无论 α 如何变化，模型的准确率都相对较低，说明基于图片分析得到的情感准确率较低，准确率较高的情况集中在 α 和 β 值都较小的区域。同样在图 6 中，当 β 的取值为 0.4 时，模型的 F1 值较小情感分析效果较差。当 α 和 β 值都较小时，模型效果与极端取值 0.1 或 0.4 相比，整体效果良好。说明在最终的情感极性分析中，图文多模态特征的情感分析结果应占据主导地位，文本和图片的情感分类结果作为辅助模块减小模型偏差。

从图 5 和图 6 中可以直观的看出，当 α，β 的取值均为 0.2 时，模型的准确率和 F1 值都达到最高值。在此系数下，文本特征情感结果和图片特征情感结果，对 SoD-ViLT 模型最终判定的情感极性贡献相同，基于图文多模态特征得到的情感结果贡献值最大。

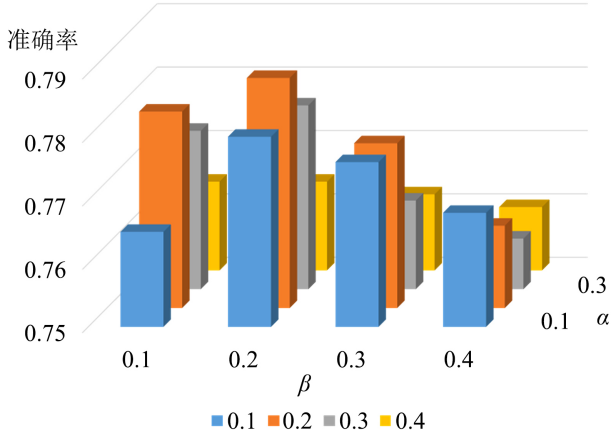

Figure 5. Accuracy at different weighting factors
图 5. 不同权重系数模型准确率

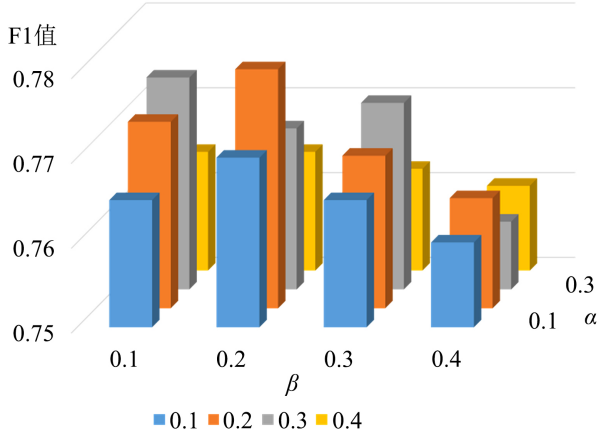

Figure 6. F1 Value at different weighting factors
图 6. 不同权重系数模型 F1 值

5. 结论

本文提出了一种基于 ViLT 的社交媒体领域图文情感分析方法 SoD-ViLT。考虑到社交媒体数据的特点，在文本特征提取模块，引入 BRERTweet 初始化权重系数。多模态特征融合阶段采用单流结构，将文本特征与图像特征进行拼接，送入同一个 Transformer 模块，得到基于图文多模态分析的情感结果。利用文本与图像自身的特征得到基于文本特征的情感分析结果和基于图片特征的情感分析结果，对三种情感分析结果使用加权融合策略确定最终的情感极性。SoD-ViLT 在 MVSA-Single 数据集和 MVSA-Multiple 数据集上的实验结果相较于 ViLT 模型，准确率分别提高了 2.73%和 2.52%，F1 值分别提高了 2.11%和 2.77%。在图文方面级情感分析任务中，SoD-ViLT 相较于 ViLT 模型，在 Twitter-15 和 Twitter-17 两个实验数据集上，准确率和 F1 值都有大幅提高。

参考文献

[1] Chen, T., SalahEldeen, H., He, X., *et al*. (2015) VELDA: Relating an Image Tweet's Text and Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, **29**, 30-36. https://doi.org/10.1609/aaai.v29i1.9168

[2] Kim, W., Son, B. and Kim, I. (2021) ViLT: Vision-and-Language Transformer without Convolution or Region Supervision. *International Conference on Machine Learning*, 18-24 July 2021, 5583-5594.

[3] Xu, N., Mao, W. and Chen, G. (2018) A Co-Memory Network for Multimodal Sentiment Analysis. *Proceedings of the* 41*st International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ann Arbor, 8-12 July 2018, 929-932. https://doi.org/10.1145/3209978.3210093

[4] Zadeh, A., Liang, P., Poria, S., *et al*. (2018) Multi-Attention Recurrent Network for Human Communication Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 5642-5649. https://doi.org/10.1609/aaai.v32i1.12024

[5] Cai, G. and Xia, B. (2015) Convolutional Neural Networks for Multimedia Sentiment Analysis. 4*th CCF Conference, NLPCC* 2015, Nanchang, 9-13 October 2015, 159-167. https://doi.org/10.1007/978-3-319-25207-0_14

[6] Huang, F., Zhang, X., Zhao, Z., *et al*. (2019) Image-Text Sentiment Analysis via Deep Multimodal Attentive Fusion. *Knowledge-Based Systems*, **167**, 26-37. https://doi.org/10.1016/j.knosys.2019.01.019

[7] You, Q., Cao, L., Jin, H., *et al*. (2016) Robust Visual-Textual Sentiment Analysis: When Attention Meets Tree-Structured Recursive Neural Networks. *Proceedings of the ACM International Conference on Multimedia*, Amsterdam, 15-19 October 2016, 1008-1017. https://doi.org/10.1145/2964284.2964288

[8] 凌海彬, 缪裕青, 张万桢, 等. 多特征融合的图文微博情感分析[J]. 计算机应用研究, 2020, 37(7): 1935-1939, 1951.

[9] 蔡宇扬, 蒙祖强. 基于模态信息交互的多模态情感分析[J]. 计算机应用研究, 2023, 40(9): 2603-2608.

[10] Xu, N., Mao, W. and Chen, G. (2019) Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 371-378. https://doi.org/10.1609/aaai.v33i01.3301371

[11] Yu, J. and Jiang, J. (2019) Adapting BERT for Target-Oriented Multimodal Sentiment Classification. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, 10-16 August 2019, 5408-5414. https://doi.org/10.24963/ijcai.2019/751

[12] Khan, Z. and Fu, Y. (2021) Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. *Proceedings of the* 29*th ACM International Conference on Multimedia*, 20-24 October 2021, 3034-3042. https://doi.org/10.1145/3474085.3475692

[13] Devlin, J., Chang, M., Lee, K., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, 2-7 June 2019, 4171-4186.

[14] Nguyen, D., Vu, T. and Nguyen, A. (2020) BERTweet: A Pre-Trained Language Model for English Tweets. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing*: *System Demonstrations*, October 2020, 9-14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

[15] Yang, X., Feng, S., Wang, D., *et al*. (2020) Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Transactions on Multimedia*, **23**, 4014-4026. https://doi.org/10.1109/TMM.2020.3035277

[16] Zhang, Q., *et al*. (2018) Adaptive Co-Attention Network for Named Entity Recognition in Tweets. *Proceedings of the Association for the Advance of Artificial Intelligence*, **32**, 5674-5681. https://doi.org/10.1609/aaai.v32i1.11962

[17] Thelwall, M., Buckley, K., Paltoglou, G., *et al*. (2010) Sentiment Strength Detection in Short Informal Text. *Journal of the Association for Information Science and Technology*, **61**, 2544-2558. https://doi.org/10.1002/asi.21416

[18] Cao, D., Ji, R., Lin, D., *et al*. (2016) A Cross-Media Public Sentiment Analysis System for Microblog. *Multimedia Systems*, **22**, 479-486. https://doi.org/10.1007/s00530-014-0407-8

[19] 郭可心, 张宇翔. 基于多层次空间注意力的图文评论情感分析方法[J]. 计算机应用, 2021, 41(10): 2835-2841.

[20] Ma, D., Li, S., Zhang, X., *et al*. (2017) Interactive Attention Networks for Aspect-Level Sentiment Classification. *Proceedings of the International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 4068-4074. https://doi.org/10.24963/ijcai.2017/568

[21] Chen, P., Sun, Z., Bing, L., *et al*. (2017) Recurrent Attention Network on Memory for Aspect Sentiment Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Copenhagen, September 2017, 452-461. https://doi.org/10.18653/v1/D17-1047

[22] Fan, F., Feng, Y. and Zhao, D. (2018) Multi-Grained Attention Network for Aspect-Level Sentiment Classification. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, October-November 2018, 3433-3442. https://doi.org/10.18653/v1/D18-1380