scientific reports

Check for updates

OPEN

基于监督渐进机器学习的句子级情感分析

苏静、陈群、王艳艳、张丽君、潘伟、李占怀

句子级情感分析（SLSA）旨在识别给定句子中传达的整体情感极性。 SLSA 最先进的性能是通过深度学习模型实现的。然而，根据 i.i.d（独立同分布）假设，这些深度学习模型的性能可能会在实际场景中表现不佳，因为训练数据和目标数据的分布几乎肯定在某种程度上有所不同。在本文中，我们为 SLSA 提出了一种基于渐进式机器学习 (GML) 的非独立同分布范式的监督解决方案。它从一些标记的观察开始，并通过迭代知识传递按照硬度增加的顺序逐渐标记目标实例。它利用标记样本进行有监督的深度特征提取，并根据提取的特征构建因子图以实现渐进式知识传递。具体来说，它使用极性分类器来检测嵌入空间中的近邻之间的极性相似性，并使用单独的二元语义网络来提取任意实例之间的隐式极性关系。我们对基准数据集的广泛实验表明，所提出的方法在所有基准数据集上实现了最先进的性能。我们的工作清楚地表明，通过利用 DNN 进行特征提取，GML 可以轻松超越纯 DNN 解决方案。

句子级情感分析（SLSA）旨在分析句子中表达的观点和情感。与方面级情感分析 (ALSA) 不同，ALSA 会分析针对特定方面表达的局部情感极性，而 SLSA 需要检测整个句子的总体情感方向。在实践中，SLSA 在评论由具有任意主题的简洁孤立句子表示的场景中非常有价值，需要对句子级别的情感进行整体分析。例如，在电子商务中，平台（例如淘宝网和预订网）可以通过了解消费者的偏好和购买体验来优化营销策略；产品制造商，例如智能手机或计算机生产商，可以根据客户反馈改进产品设计。在另一个应用中，社交媒体平台（例如 Twitter 和 Facebook）通常通过 SLSA 分析人们的评论和帖子，以深入了解公众舆论和社会趋势。

SLSA 最先进的性能是通过各种 DNN 模型实现的。特别是，过去几年的经验表明，通过大规模语料库上的语义学习，预训练模型（例如 BERT 、 RoBERTa 和 XLNet ）可以自动捕获隐含的情感特征，从而有效提高SLSA。然而，这些 DNN 模型的有效性取决于 i.i.d（独立同分布）假设；但在实际场景中，可能没有足够的标记训练数据，即使提供了足够的训练数据，训练数据和目标数据的分布也几乎肯定存在一定程度的不同。我们通过如图 1 所示的运行示例来说明 SLSA 的挑战，其中我们通过颜色深度指示单词的情感极性。在S中，第一部分表示正极性，而第二部分表示负极性。因此，其整体极性为负。不幸的是，BERT 模型将其极性错误地识别为正。在S中，BERT模型未能检测到“not”和“long”组合的正极性。

为了减轻训练数据和目标数据之间分布不一致所造成的限制，本文基于最近提出的渐进机器学习的非独立同分布范式，提出了一种 SLSA 的监督方法。一般来说，GML 从一些简单的实例开始，然后通过标记和未标记实例之间的知识传递逐渐标记更具挑战性的实例。从技术上讲，GML通过因子图中的迭代因子推理来实现渐进式知识传递。值得注意的是，与独立同分布学习方法（例如深度学习）不同，独立同分布学习方法为目标工作负载中的所有实例训练单个统一模型，

西北工业大学计算机学院, 陕西 西安 710072邮箱：sujing@
西北工业大学邮箱

ID	真实标签	预测标签	词义
S_0	消极的	积极的	一个仍然工作正常，另一个一天后就退出了。
S_1	积极的	消极的	学习如何使用它不会花费很长时间。
S_2	消极的	积极的	键盘大小合适，但电源开/关键很小，很难按。

图 1. CR 数据集中 SLSA 的说明性示例：BERT 模型对所有三个句子都做出了错误预测。

GML 根据不断变化的证据观察逐渐了解每个实例的标签状态。通过逐步学习，GML 可以有效地桥接标记训练数据和未标记目标数据之间的分布对齐。 GML 已成功应用于方面级情感分析 (ALSA) 以及实体解析的任务。即使不利用标记的训练数据，现有的无监督 GML 解决方案也可以实现与有监督 DNN 模型相比具有竞争力的性能。然而，这些无监督解决方案的性能仍然受到知识传递不准确和不足的限制。例如，现有的用于方面级情感分析的 GML解决方案主要利用情感词典和由话语结构指示的显式极性关系来实现情感知识传达。一方面，情感词典可能不完整，情感词的实际极性在不同的句子上下文中可能会有所不同；另一方面，自然语言语料库中明确的极性关系通常很稀疏。因此，它们作为情感知识传递媒介的功效是有限的。

我们提出的 SLSA GML 解决方案旨在有效利用标记的训练数据来增强渐进学习。具体来说，它利用二元极性关系（这是最直接的知识传递方式）来实现监督渐进学习。由于人们普遍认为基于 BERT 的预训练模型可以比手动制作的特征（例如情感词典）更准确地捕获情感特征，因此我们利用标记的训练数据通过基于 BERT 的模型提取情感特征。与现有的 DNN 模型类似，它训练句子级极性分类器，使得具有相似极性的句子可以在深度嵌入空间的局部邻域内聚类。为了使知识传递超越局部邻域，我们还单独训练语义网络来提取两个任意句子之间的隐式极性关系。然后将所有提取的特征建模为因子图中的二元因子以实现逐步学习。我们通过图2中的例子说明了渐进推理的过程。渐进知识传递应该是通过二元因素来实现的。在该示例中，给定证据观察和二元相似性因子，t、t 和 t 的标签随后可以被推理为负。

本文的主要贡献可概括如下：

- 我们为 SLSA 提出了一种有监督的 GML 解决方案，它可以有效地利用标记的训练数据来增强渐进学习；
- 我们提出了两种类型的 DNN 模型来捕获隐含的情感特征，并将它们建模为因子图中的二元因子，以实现 SLSA 的监督知识传递；
- 我们通过比较研究实证验证了所提出的解决方案在实际基准工作负载上的有效性。我们的大量实验表明，它在所有测试工作负载中始终达到最先进的性能。

本文的其余部分组织如下。 “相关工作”部分讨论了相关工作。 “预备知识”定义了 SLSA 的任务并介绍了 GML 框架。 “有监督的 GML 解决方案”提出了建议

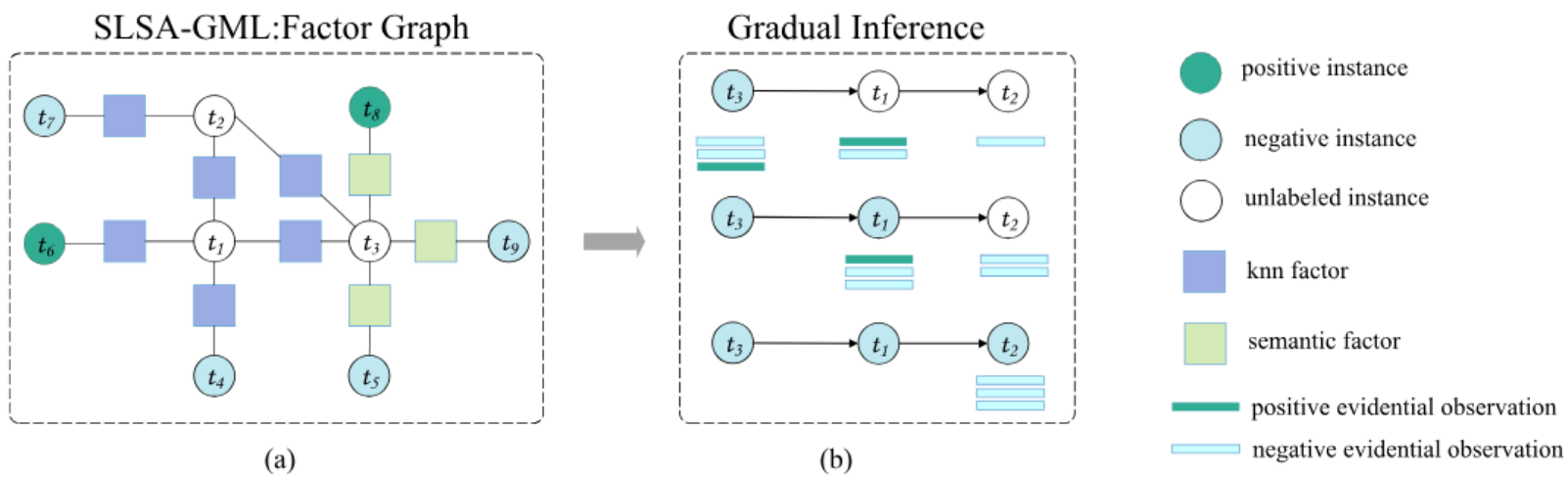

图2.渐进学习的说明性示例：1）子图（a）表示将未标记和标记样本与两种类型的关系特征（knn特征和语义特征）连接起来的因子图； 2）子图（b）说明了基于未标记样本的证据观察的逐步推理过程。 t、t 和 tare 的标签随后被推断为负数。

解决方案。“实证评估”对所提出的解决方案进行实证评估。在“结论”中，我们对本文进行了总结，并对未来的工作进行了一些思考。

相关工作

文献中已经在不同粒度（例如文档级、句子级和方面级）上广泛研究了情感分析。在文档级别，目标是检测整个评论的情感极性，该评论可能由多个句子组成。句子级情感分析旨在检测单个句子中表达的一般极性。代表最精细的粒度，方面级情感分析需要识别句子中实体的某些方面所表达的极性。值得注意的是，一个句子可能表达了句子中不同方面的冲突极性。不同粒度的最先进的情感分析解决方案是建立在 DNN 模型的基础上的。例如，对于文档级情感分析，DNN 模型包括 CSNN 、 AttBiLSTM-2DCNN 、 CNN-BiLSTM 、 SR-LSTM 和 BAE ； 对于方面级别的情感分析，最新的 DNN 模型包括 LCF-BERT、PTM 和 RGAT，它们都是预训练 BERT 模型的变体。本文重点关注句子层面的情感分析。在本节的其余部分中，我们从句子级情感分析和渐进式机器学习的正交角度回顾相关工作。

句子级情感分析。

SLSA 的早期工作主要集中在为 SVM 分类器提取不同的情感提示（例如 n-gram、词典、pos 和手工规则）。不幸的是，这些特征要么稀疏，仅覆盖几个句子，要么不高度准确。深度神经网络的进步使得特征工程对于许多自然语言处理任务来说变得不必要，特别是包括情感分析。最近，人们提出了各种基于注意力的神经网络来更准确地捕获细粒度的情感特征。不幸的是，这些模型不够深入，因此对极性检测的功效有限。

最近，SLSA 的研究经历了向大型预训练语言模型（例如 BERT、RoBERTa 和 XLNet）的重大转变。一些研究人员研究了如何将传统语言特征（例如词性、语法依赖树和知识库）集成到预训练模型中以提高性能。其他研究人员关注如何基于标准变压器结构设计新的情感分析网络。通常，他们将 BERT 模型的输出馈送到新网络，将原始预训练模型的参数重新加载到新网络。随后，提出了一些新的预训练建议，以减轻新网络结构和预训练模型之间的不匹配。例如，SentiLARE 将情感分数编码为输入嵌入的一部分，并对 yelp 数据集进行后预训练以获得自己的预训练模型。 Entailment 的工作修改了预训练过程，生成了新的预训练模型 SKEP_ERNIE_2.0_LARGE_EN 。

此外，为了更好地使预训练模型适应下游任务，一些研究人员提出设计新的预训练任务。例如，SentiBERT 的工作设计了特定的预训练任务来指导模型预测短语级情感标签。 Entailment 的工作将多个 NLP 任务（包括句子级情感分析）重新表述为统一的文本蕴涵任务。值得注意的是，到目前为止，该方法在句子级情感分析上取得了最先进的性能。

值得注意的是，上述所有 SLSA 深度学习解决方案都是基于 i.i.d 学习范式构建的。对于 SLSA 的下游任务，其实际效果通常取决于足够大量的标记训练数据。然而，在实际场景中，可能没有足够的标记训练数据，即使提供了足够的训练数据，训练数据和目标数据的分布几乎肯定存在一定程度的不同。

渐进式机器学习。

渐进式机器学习（GML）的非独立同分布学习范式最初是为了实体解析任务而提出的。它可以按照硬度增加的顺序逐渐标记实例，而不需要手动标记工作。此后，GML也被应用于方面级别的情感分析任务。值得指出的是，作为一种通用范式，GML 可能适用于各种分类任务，包括本文所示的句子级情感分析。尽管现有的无监督 GML 解决方案与许多有监督方法相比可以实现有竞争力的性能，但在不利用标记训练数据的情况下，它们的性能仍然受到不准确和不足的知识传递的限制。在本文中，我们重点关注如何监督 DNN 的特征提取，并利用它们来改进 SLSA 任务的渐进学习。

预赛

在本节中，我们首先定义 SLSA 任务，然后提供 GML 框架的简要概述。

任务定义。

与往常一样，本文将 SLSA 视为二元分类问题，其中分类器需要将每个句子标记为肯定或否定。正式地，我们将 SLSA 的任务定义如下:

定义 1（句子级情感分析）给定评论语料库{r,r,r, 。 。 。 ,r} 每条评论 r 组成-一系列句子的荷兰国际集团，{s,s,s, 。 。 。 ,s}，SLSA的目标是预测每个句子的标签，其中标签可以是正数（标签 = 1 标签 = 0

通用 GML 框架。

如图3所示，通用的GML框架由以下部分组成三个基本步骤:

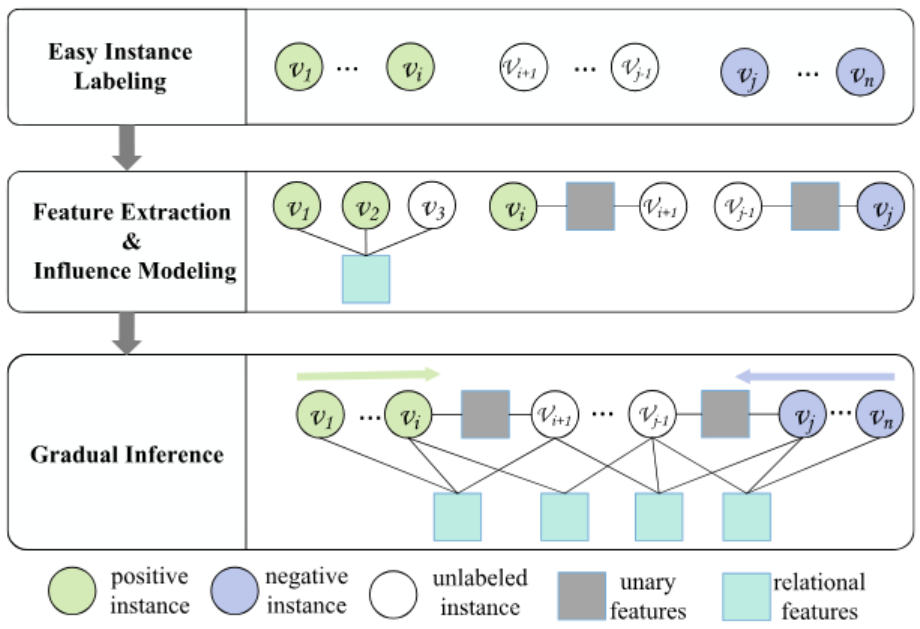

图 3. 通用 GML 框架。

简单的实例标记。 渐进式机器学习从简单实例的标签观察开始。在无监督设置中，通常可以根据专家指定的规则或无监督学习来执行简单的实例标记。例如，可以观察到，如果一个实例非常接近聚类中心，那么它通常只有很小的机会被错误分类。因此，它可以被认为是一个简单的实例并自动标记。

对于方面级别的情感分析，已经表明，如果一个句子包含一些强积极的内容（res.
否定）情感词，但没有否定、对比和假设连接词，可以可靠地推理为肯定（res.否定）。在本文中，我们研究了监督环境中的句子级情感分析，其中一些标记的训练数据应该是可用的。这些带有真实标签的训练实例自然可以作为初始简单实例。

特征提取和影响建模。 在 GML 中，特征充当标记实例和未标记实例之间知识传递的媒介。通常需要提取各种各样的特征来捕获不同的信息。对于每种类型的特征，此步骤还需要对其对标签状态的影响进行建模。不同的应用程序需要不同的功能，这是很常见的。在我们之前关于用于方面级情感分析的无监督 GML 的工作中，我们提取了由话语结构指示的情感词和明确的极性关系，以促进知识传递。不幸的是，对于句子级情感分析，句子之间很少存在极性关系提示，并且情感词通常不完整且不准确。因此，我们建议使用 DNN 来提取隐含的情感特征。

逐渐推理。

此步骤逐渐标记工作负载中难度不断增加的实例。 GML 通过对由标记和未标记实例及其共同特征组成的因子图进行迭代因子推理来实现渐进学习。在每次迭代中，它通常会以最高程度的证据确定性来标记未标记的实例。

Algorithm 1 Scalable Gradual Inference Algorithm

while *there exists any unlabeled variable in* G **do**
  $V' \leftarrow$ all the unlabeled variables in G;
  for $v \in V'$ **do**
    Measure the evidential support of v in G;
  Select top-m unlabeled variables with the most evidential support (denoted by V_m);
  for $v \in V_m$ **do**
    Approximately rank the entropy of v in V_m;
  Select top-k most promising variables in terms of entropy in V_m (denoted by V_k);
  for $v \in V_k$ **do**
    Compute the probability of v in G by factor graph inference over a subgraph of G;
  Label the variable with the minimal entropy in V_k;

形式上，假设类标签的总数表示为 {L ，L，。。。，L。给出一个推论变量 v，GML 通过熵的倒数来衡量其证据确定性，如下所示

$$E(v) = H(v) = \sum_{L} P(v) \frac{1}{P(v)} \log \frac{1}{P(v)}, \quad (1)$$

其中E(v)和H(v)分别表示v的证据确定性和熵，并且 $P(v)$ v 具有标签的推断概率 L

。值得注意的是，在逐步推理的过程中，当前迭代中新标记的实例将作为后续迭代中的证据观察。

在实践中，GML通常通过可扩展的渐进推理来实现，它最初是针对实体解析任务而提出的。我们概述了可扩展渐进推理的一般过程，与算法 1 中提出的相同。它由以下三个步骤组成：（1）证据支持的测量；(2)熵的近似排序；(3)因子子图推理。给定一个因子图 G，它首先选择 G 中最具证据支持的前 m 个未标记变量作为概率推理的候选变量。为了减少因子图推理的调用频率，它然后通过有效的算法对 m 个候选者进行近似熵估计，并仅选择其中最有希望的前 k 个变量进行因子图推理。最后，通过因子子图推断来推断所选 k 个变量的概率。

受监督的 GML 解决方案

监督解决方案直接使用训练数据中的标记示例作为简单实例。正如简介中提到的，它利用两种类型的 DNN 模型进行知识传递：极性分类器，用于提取极性敏感向量表示，以检测深度嵌入空间中近邻之间的极性相似性；以及语义深度网络，它可以检测两个任意句子之间的相似和相反极性关系。然后，我们的解决方案构建一个由一元和二元因子组成的因子图，以实现逐步学习。

我们通过如图 4 所示的示例来说明所提出的解决方案。在该示例中，存在三个未标记的实例：t、t 和 t。子图（a）和（b）分别显示了它们基于邻域的相似性特征和语义关系特征，子图（c）显示了构建的因子图。在本节的其余部分中，我们首先介绍 DNN 模型来提取极性关系特征，然后描述如何将它们建模为因子图中的因子以促进逐步学习。为了简单起见，我们在表 1 中总结了常用的符号。

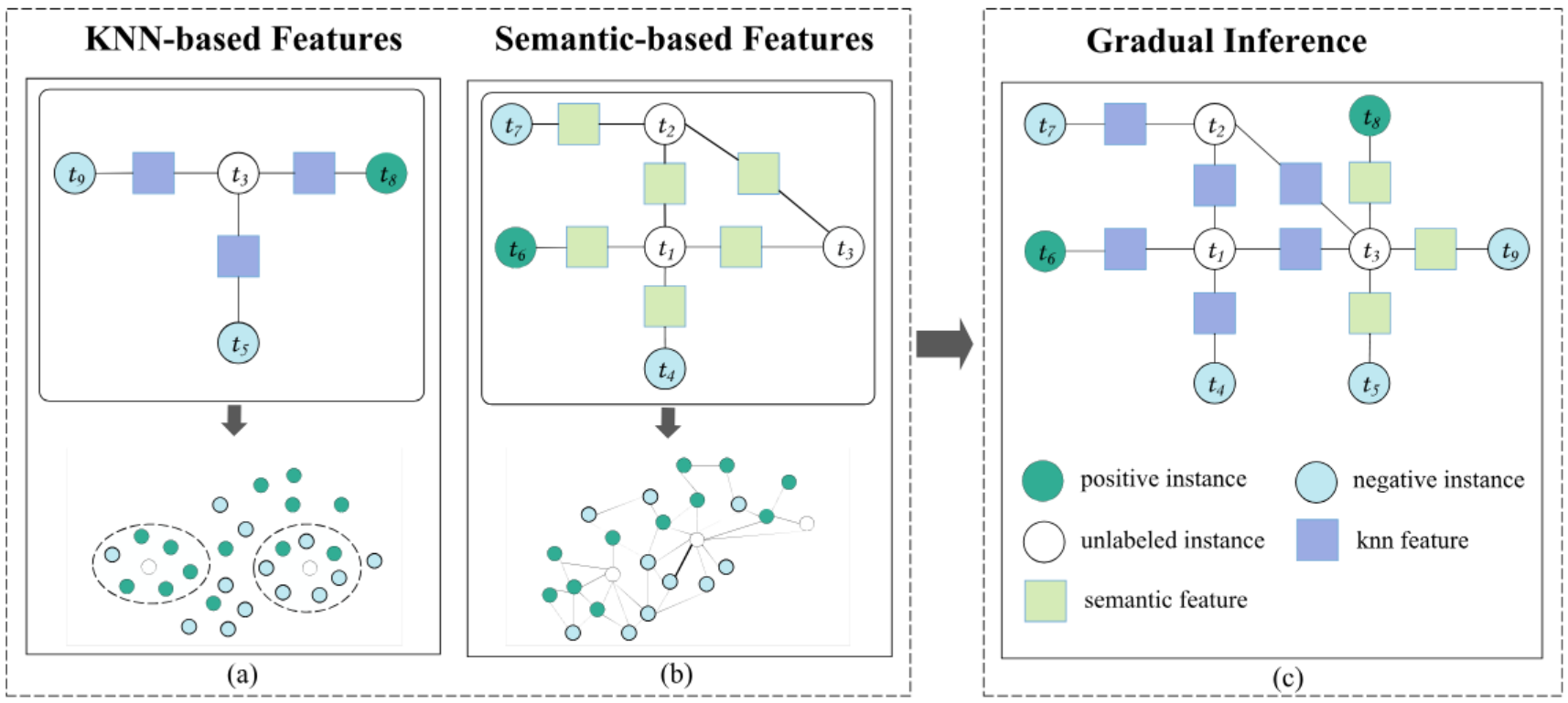

图 4. SLSA 的监督 GML 解决方案说明：它提取两种类型的极性关系特征，一种基于局部邻域，另一种基于语义深层网络。

符号	描述
v	情感向量
h 隐藏状态向量	
h	情感感知隐藏状态向量
w 情感注意力权重向量	
w	词 i 的情感注意力权重
w	d维情感权重（例如Evaluation的权重）
v	单词 i 的第 d 维情感值（例如评估中单词 i 的情感得分）

表 1. 经常使用的符号。

特征提取：极性关系。

由于最近的工作表明，对手工制作的情感知识（例如情感词典）进行编码可以有效增强深度 DNN 的极性分类训练，因此我们的解决方案在基于 EFL 的 DNN 模型上添加了新的情感关注分支，以生成极性敏感的嵌入空间。

极性感知嵌入的相似性。

情感注意力的新分支由情感学习层和池化层组成，用于反映情感词典所指示的每个单词的明确情感极性。在情感学习层，我们使用两个情感词典 EPA（评估、效力和活动）和 VAD（价、唤醒和优势）构建情感向量，这两个词典都通过连续数值的三个独立维度来衡量情感取向。具体来说，我们将每个单词的两个单词级向量连接成一个 6 维情感向量 v，其中六个维度分别对应于评估、效力、活动、效价、唤醒和优势。在模型中，编码器层的输出是一个隐藏状态向量，表示为

$h \in R$，然后被馈送到池化层，其中b表示批量大小，m表示序列的最大长度，e表示嵌入维度。在将 h 传送到池化层之前，我们将 h 转换为情感感知隐藏状态的新向量，表示为 h。具体来说，我们通过情感维度值的加权和来衡量每个情感词的注意力权重，如下所示：

$$w = 1 + \sum_{d=0}^{5} (w \times v), \tag{2}$$

其中 w 表示情感词的注意力权重，w 表示第 d 个情感维度权重，and v 表示该词在第 d 个情感维度上的情感值，请注意，值 w 表示六个维度的权重（即评价、效能、活性、效价、唤醒度和支配力）；在我们的实现中，我们按照建议将它们的值设置为 [0.2, 0.2, 0.3, 0.3, 0.2, 0.2] 。的价值 v 表示单词在第 d 个情感维度的情感值，可以直接从 EPA 和 VAD 词典中提取。还值得注意的是，EPA词典的维度值域是[0,1]，而VAD的维度值域是[−5,5]。因此，我们使用映射函数将EPA和VAD的域统一在[0,1]处。根据设置 w and v，值域 w 介于 1 和 2.4 之间。如果缺少一个词在情感词典中，我们将其注意力权重设置为 1，或者 $w = 1$，有效地忽略了词典的影响。

接下来，我们将句子中所有单词的注意力权重连接起来，得到其情感注意力权重向量w，如下所示：

$$w = [w^1, w, 。 。 。, w] \tag{3}$$

在哪里 $w \in R$。然后，我们计算新的情感感知隐藏状态向量 h 通过对原始数据进行加权隐藏状态向量 h 和 w 如下：

$$h = w \times h. \tag{4}$$

在最终的分类层中，我们融合两个分支的特征，如下所示：

$$l = l^n + \theta \times l^\xi, \tag{5}$$

其中land l分别表示表示学习分支和情感注意力分支产生的损失，θ 表示平衡两个分支贡献的惩罚权重参数。由于情感注意分支应该补充主要表示学习分支，因此我们建议设置 θ 小于0.5。在我们的实现中，我们设置 $\theta = 0.3$.

具体来说，我们有

$$l = y \text{ 对数 } y + (1 - y) \text{ 日志}(1 - y), \tag{6}$$

and

$$l = y \text{ 对数 } y + (1 - y) \text{ 日志}(1 - \hat{y}). \tag{7}$$

我们使用标记的训练数据对极性分类模型进行微调，如图 5 所示，然后利用所得的向量表示（最后一层嵌入）进行极性相似性检测。在实现中，我们基于最先进的 EFL 模型构建了极性分类的 DNN。对于目标工作负载中的每个未标记句子，我们从标记实例和未标记实例中提取其 k 最近邻。我们使用余弦距离来衡量相似度。 k的值通常设置为较小的数字以保证提取的关系的准确性。此外，我们使用阈值（例如，在我们的实验中为 0.001）来过滤掉嵌入空间中不够接近的最近邻居。我们的实验证明，只要 k 值设置在合理的范围内（1 到 9 之间），监督 GML 的性能对于 k 值来说是稳健的。

通过语义深层网络的相似/相反关系。 语义深度网络建立在 Transformer 架构之上，旨在检测两个任意句子之间的极性关系。 Transformer 的骨干是由多个多头自注意力层组成的编码器。每层具有相同的网络结构但参数权重不同。众所周知，在 Transformer 中，除了最后一个隐藏层之外，其他层也包含情感信息。因此，我们添加一个自注意力层来聚合变压器最后五层中存在的信息，并使用超级特征向量来捕获最后一层之外的额外情感特征。

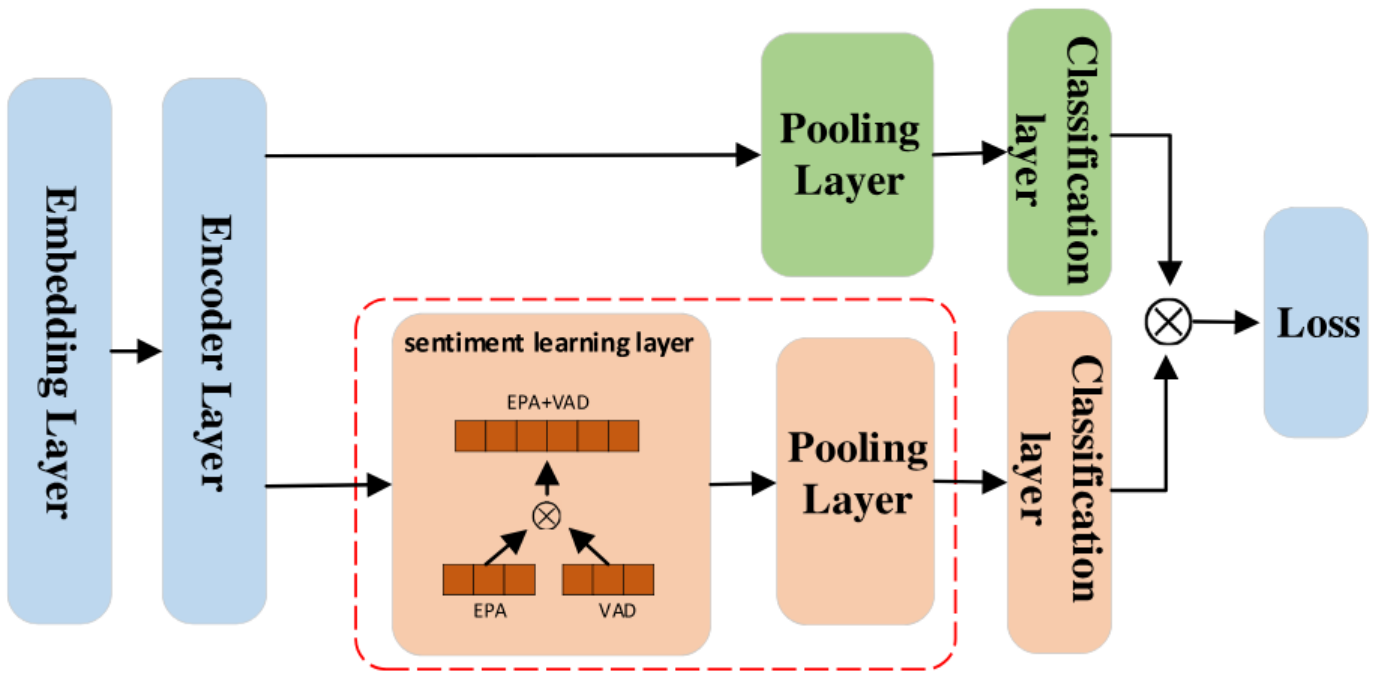

图 5 基于 EFL 的极性分类模型的结构： 1）红色虚线圈出的框代表新添加的情感注意力编码情感知识的分支；2）橙色块代表新添加的情感学习层的组件，绿色块是原始EFL模型中存在的组件，蓝色块是这两个分支之间的共享组件。

具体来说，如图6所示，语义深度网络的结构可以用以下三个方程表示:

$$h = h \oplus h \oplus h \oplus h \oplus h, \quad (8)$$

$$h = f(h), \quad (9)$$

$$h = w \times h + b. \quad (10)$$

在等式中。(8)表示串联操作，$h \in R$ 表示前第i个隐藏层的向量输出层，其中 $i \in \{19, 20, 21, 22, 23\}$ 和 $h \in R$ 表示序列的最大长度，e表示嵌入维度，$h \in R$ 表示连接后的超向量，$h \in R$.

方程（9）利用了多头自注意力函数 f 在超级向量上学习更多不同的特征。

最后，等式。(10) 使用线性函数来映射尺寸大小 $b \times m \times 5e$ 到原始尺寸 of $b \times m \times e$ 对于后续层，其中 $w \in R^{e}$ 表示权重矩阵，b 表示偏差，两者这些应该通过培训来学习。

设超向量为 h ，多头自注意力函数的第一步是通过线性变换将原始输入向量映射为查询向量、键向量和值向量，如下所示:

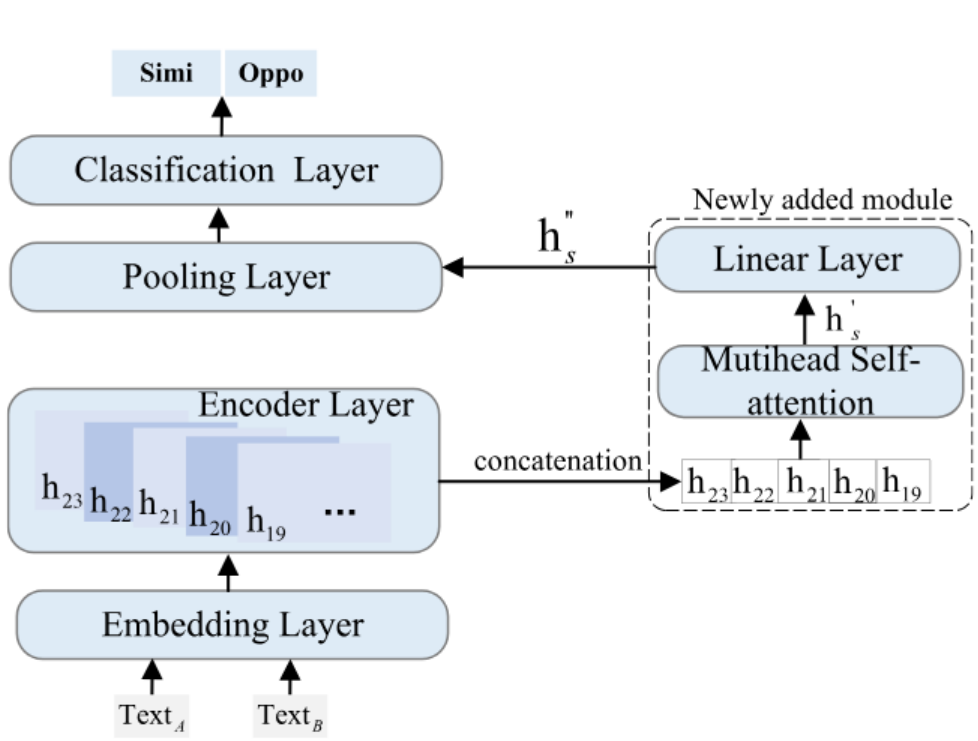

图 6. 用于任意极性关系提取的语义深度网络的结构：右侧的虚线框表示我们新添加的模块。

$$q = w \times h + b,\tag{11}$$

$$k = w \times h + b,\tag{12}$$

$$v = w \times h + b,\tag{13}$$

其中 qkv 分别表示查询向量、键向量和值向量，www 矩阵的大小为 $Rbbb$

然后，它利用 softmax 函数将查询向量和关键向量转换为注意力概率，如下所示：

$$a = \text{软最大}\left(\frac{q \times k}{\sqrt{\gamma}}\right)\tag{14}$$

其中 $a \in R$。最后，它会乘以 a with v 获得组合的上下文信息单词特征 h as 如下：

$$h = a \times v\tag{15}$$

然后，它就转变了 hh 根据方程式（10）。其后续处理与传统类似变压器架构。

对于 SLSA，我们基于经过训练的语义深度网络构建标记和未标记句子之间的极性关系。在训练阶段，我们从每个标记句子的训练数据中随机提取r个标记句子来微调语义网络。然后，在特征提取阶段，我们从目标工作负载中每个未标记句子的标记训练数据中随机提取 r 个句子，并基于语义网络构建其与它们的关系。我们的实验表明，只要 r 的值设置在合理的范围内，监督 GML 的性能就非常稳健（$3 \leq r \leq 8$

二元关系的因子建模。

渐进式机器学习的因子图由证据变量、推理变量和因子组成。在 SLSA 中，一个变量对应一个句子，一个因子定义两个变量之间的二元关系。在GML的过程中，需要逐步推断推理变量的标签。变量的标签一旦推断就保持不变。说明性示例的因子图如图4所示。

在监督环境中，所有标记的训练数据都作为简单实例。目标工作负载中句子的标签需要通过提取的二元关系通过知识传递逐步推断出来。具体来说，我们定义基于 KNN 的相似关系的二元因子，f, as

$$\phi(v, v) = \begin{cases} e & \text{if } v = v; \\ 1 & \text{否则;} \end{cases}\tag{16}$$

在哪里 vv 表示共享基于 KNN 的相似关系特征的两个变量 fw 的重量 f。类似地，我们定义两个变量之间语义关系的二元因子，f

$$\phi(v, v) = \begin{cases} e & \text{if } v = v; \\ 1 & \text{否则}; \end{cases}\tag{17}$$

在哪里 v and v 表示共享语义关系特征的两个变量 f，和 w 表示重量 f.

在我们实施可扩展的渐进推理时，相同类型的因素应该具有相同的权重。最初，相似性因素（无论是基于 KNN 的因素还是语义因素）的权重被设置为正（例如，在我们的实验中为 1），而相反的语义因素的权重被设置为负（例如，在我们的实验中为 - 1）。实验）。值得注意的是，三个参数的权重将根据推理过程中的证据观察不断学习。

在可扩展渐进推理的计算复杂度上，SLSA 的分析结果与我们之前在 ALSA 上的工作所代表的结果基本相同。具体来说，在每次标记迭代中，证据支持测量的计算复杂度可以表示为 O($n \times n$)，其中n 表示提取的特征总数(例如，熵估计) 表示提取的每个句子的近似熵估计的计算复杂度可以 n 表示为 O($m \times n$)，其中 m 表示为算法 1 中指定的近似熵估计选择的候选变量的数量；因子子图构建的计算复杂度可以表示为 O($k \times n$)，其中 k 表示为算法 1 中指定的因子图推断选择的候选变量的数量

实证评估

在本节中，我们通过比较研究来实证评估所提出的解决方案的性能。“实验设置”描述了实验设置。“比较评价”指*表示比较评价结果。“敏感性评估”评估所提出的解决方案相对算法参数的性能敏感性。“对其他分类任务的可扩展性的讨论”*讨论了所提出的方法对其他分类任务的可扩展性。

实验设置。

为了进行比较评估，我们使用电影评论（MR）、客户评论（CR）、Twitter2013和斯坦福情绪树库（SST）的基准数据集。 MR和SST都是电影评论集合，CR包含电子产品的客户评论，而Twitter2013包含微博评论，这些评论通常比电影和产品评论短。测试数据集的详细统计数据如表2所示。

由于最近提出的基于预训练语言模型的 DNN 解决方案已被经验证明优于早期的提案，因此我们将所提出的解决方案（用 GML 表示）与以下最先进的模型进行比较：

- 英语英语。将类标签转换为辅助句子，它是一个统一的模型，可以将多个 NLP 任务建模为文本蕴涵任务。
- 森蒂拉尔。作为一种语言表示模型，它将词级语言知识（包括词性标签和情感极性）引入到预训练模型中，并使用标签感知的掩码语言模型来构建知识感知的语言表示。
- 罗伯特·大号。它有目的地删除下一个句子预测目标并动态改变掩蔽模式以提高下游任务的性能。
- XLNet-大型。它基于广义自回归预训练模型，该模型可以通过最大化分解顺序的所有排列的预期可能性来学习双向上下文。
- 罗伯特塔基地。它是 RoBERTa-Large 的简单版本，只有 12 个隐藏层。
- XLNet-基地。它是 XLNet-Large 的简单版本，只有 12 个隐藏层。
- 斯伯特。它使用连体和三元组网络结构来导出语义上有意义的句子嵌入，以进行情感极性检测。
- 活动星系核。它将统计信息与语义表示相结合，以训练用于情感分析的鲁棒分类器。
- 双CL。它是最近提出的情感分析框架，可以在同一嵌入空间中同时学习输入样本的特征和分类器的参数。

我们基于开源 GML 推理引擎 (https://github.com/gml-explore/gml) 实现了所提出的 GML 解决方案。我们的实现使用情感感知 EFL 模型作为基线极性分类器，并利用外部情感知识来提取基于 KNN 的相似关系。它使用改进的 RoBERTa-Large 模型来提取任意两个句子之间相似和相反的语义关系。具体来说，RoBERTa-Large 模型由 16 个头和 24 层组成，隐藏层大小为 1024。我们的实现将 dropout 概率保持在 0.1，并将 epoch 数设置为 3。它将初始学习率设置为 $2e^{-5}$。为了训练语义深度网络，我们为每个标记示例生成 6 个语义关系（3 个具有相似标签，3 个具有相反标签）。对于GML因子图构建，我们为每个未标记句子在训练集中随机选择6个标记示例，并使用训练后的二元语义模型来预测它们的极性关系。

像往常一样，我们通过准确度和宏 F1 指标来衡量不同解决方案的性能。所有报告的结果均为 5 次运行的平均值。我们报告平均值和标准差 (STD)。所有对比实验均在同一台机器上进行，该机器运行 Ubuntu 16.04 操作系统，并具有 NVIDIA GeForce RTX 3090 GPU、128 GB 内存和 2 TB 固态硬盘。

比较评价。

详细的评估结果如表 3 所示。可以看出，GML 在所有测试工作负载中在准确度和 Macro-F1 方面始终达到了最先进的性能。具体而言，就准确率而言，GML 在 CR、MR、SST 和 Twitter2013 上分别优于现有表现最好的 EFL 1.68%、0.89%、1.76%、1.58%。同样，就宏观F1而言，GML的表现分别优于EFL 1.18%、0.81%、1.7%、1.11%。在准确性方面，GML 在 CR 上比现有最先进的技术高约 1.6%，在 MR 上比现有技术高 0.7%，在 SST 上比现有技术高 0.6%，在 Twitter2013 上比现有最先进技术高 0.58%。就 Macro-F1 而言，四个测试工作负载相对于现有技术的改进幅度分别为 1.18%、0.46%、0.55%、0.40%。值得注意的是，最近两种方法 AGN 和 DualCL 的性能与其他 DNN 模型相似，但比 GML 差。我们观察到，AGN 关注的统计特征（例如词频）对于情感分析没有太大帮助。作为一种增强方法，DualCL 通常在只有少量标记训练数据的情况下表现良好。然而，在我们的基准工作负载场景中，DualCL 的功效相当有限。通过利用最先进的 DNN 进行特征提取，非独立同分布渐进学习比独立同分布学习具有明显的优势。还可以观察到，对于 GML 和深度学习模型，

数据集	火车	验证	Test
MR	8534	1066	1050
CR	2262	754	754
推特2013	5098	915	2034
SST	6920	872	1821

表 2. 测试数据集的统计。

模型	CR		MR		SST		推特2013	
	Acc	宏-F1	Acc	宏-F1	Acc	宏-F1	Acc	宏-F1
EFL	93.94%±0.04	95.36%±0.12	92.27%±0.32	92.23%±0.22	94.51%±0.02	94.60%±0.40	93.36%±0.55	95.38%±0.45
森蒂拉尔	92.41%±0.14	94.03%±0.11	91.52%±0.10	91.42%±0.15	94.56%±0.80	94.69%±0.23	92.90%±0.26	95.02%±0.24
罗伯特·塔·拉格	93.73%±0.43	95.06%±0.27	92.13%±0.25	92.12%±0.28	95.66%±0.91	95.75%±0.64	94.36%±0.44	96.10%±0.30
XLNET-大号	93.44%±0.07	94.83%±0.04	91.31%±0.26	91.33%±0.25	95.30%±0.19	95.40%±0.73	93.80%±0.55	95.72%±0.50
罗伯特·塔基斯	93.04%±0.46	94.50%±0.31	90.23%±0.23	90.27%±0.21	94.82%±0.19	95.00%±0.13	93.51%±0.37	95.50%±0.31
XLNET-基地	92.84%±0.28	94.40%±0.15	90.09%±0.29	90.16%±0.73	93.00%±0.38	93.17%±0.38	92.51%±0.09	94.84%±0.09
SBERT	92.93%±0.10	93.55%±0.32	92.38%±0.45	92.58%±0.40	95.09%±0.80	95.22%±0.21	93.25%±0.11	94.31%±0.32
AGN	91.89%±0.13	91.24%±0.20	87.60%±0.32	87.57%±0.13	92.72%±0.20	90.94%±0.29	91.26%±0.30	91.26%±0.45
双CL	92.19%±0.28	92.68%±0.85	89.41%±0.54	89.06%±0.13	93.41%±0.10	93.58%±0.77	88.94%±0.24	89.05%±0.22
SLSA-GML	95.62%±0.21	96.54%±0.29	93.16%±0.30	93.04%±0.32	96.27%±0.12	96.30%±0.15	94.94%±0.20	96.49%±0.20

表 3. 比较评估结果：我们以粗体突出显示每个数据集的最佳结果。

不同运行的波动仍然很低（大多数情况下 STD 值 < 0.5）。由于 SLSA 的挑战得到了广泛认可，这些观察结果清楚地表明了所提出方法的有效性。

消融研究。 我们还对所提出的 GML 解决方案进行了消融研究。详细的评估结果如表 4 所示。可以看出，如果没有基于 KNN 的关系或二元语义关系，GML 在所有测试工作负载上的性能都会下降。这一观察结果清楚地表明，基于 KNN 的关系和二元语义关系是互补的：它们在 GML 中的组合建模比它们中的任何一个都取得了更好的性能。然而，也可以观察到，与 knn 关系相比，没有二元语义关系的 GML 性能下降得更明显。 knn 关系仅捕获相似性特征，而二元语义关系可以捕获相似性和相反或更多样化的关系。值得注意的是，我们的实验结果与 GML 的预期特征一致，即更多样化的特征通常可以更有效地促进知识传递。

说明性例子。 我们通过 CR 的例子说明了 GML 的有效性，如表 5 和图 7 所示。在 t 上，GML 和深度学习模型都给出了正确的标签；然而，在所有其他示例中，GML 给出了正确的标签，而深度学习模型则给出了错误的预测。在图7中，四个子图分别显示了示例构建的因子子图。可以看出，这三个相关因素，其中两个是正确预测的，而其余一个是错误预测的。然而，GML仍然正确

模型	CR		5MR		SST		推特2013	
	Acc	宏-F1	Acc	宏-F1	Acc	宏-F1	Acc	宏-F1
SLSA-GML（不带 knn)	95.36%±0.20	96.32%±0.35	93.07%±0.17	92.97%±0.72	96.16%±0.71	96.20%±0.45	94.74%±0.23	96.34%±0.85
SLSA-GML（不带语义）	94.56%±0.75	95.69%±0.25	93.06%±0.89	92.94%±0.65	95.61%±0.27	95.66%±0.55	93.90%±0.86	95.78%±0.95
SLSA-GML	95.62%±0.74	96.54%±0.06	93.16%±0.28	93.04%±0.90	96.27%±0.39	96.30%±0.43	94.94%±0.41	96.49%±0.11

表4.消融研究的评估结果：我们分别评估KNN和语义因素的功效。重要值以[粗体]显示。

#Id	#文本	地面真实标签	GML	DNN
t	这款相机非常适合热情的业余摄影师。	Pos	Pos	Pos
但我已经通过电子邮件向创意技术支持发送了有关此问题的电子邮件，并得到了及时的回复 - 如有必要，他们会为我解决。		Pos	Pos	Neg
t	这确实成功地将启动期间弹出窗口的数量从 4 个减少到 2 个。	Pos	Pos	Neg
t	我想象如果我让我的播放器保持不变（无背光），它可以在低音量下播放超过 12 小时。	Pos	Pos	Neg

表 5. CR 数据集中的 GML 功效的说明性示例； GML 列代表我们提出的 GML 解决方案的预测结果，而 DNN 列代表当前 SOTA DNN 模型的预测结果。

预测“t”的标签，因为它的大多数关系对应项都表示正极性。在 t 上也观察到了类似的结果。值得注意的是，GML 按照 t、t、t 和 t 的顺序标记这些示例
t
。由于 t 的预测标签提供了具有正确极性提示的标签，因此 t 也被正确标记为正。

敏感性评估。

我们还分别根据提取的语义关系的数量和提取的 KNN 关系的数量评估了 GML 的性能敏感性。两个参数都设置在 3 到 8 之间的范围内。详细的评估结果如表 6 所示。可以看出，GML 的性能对于这两个参数都非常稳健。这些实验结果预示着 GML 在实际场景中的适用性。

关于其他分类任务的可扩展性的讨论。

可以看出，我们提出的方法利用二元标签关系（知识传递的通用机制）来实现渐进学习。对于其他分类任务，例如方面级或文档级情感分析，甚至是更普遍的文本分类问题，由于 DNN 分类器的可用性，生成基于 KNN 的关系特征非常简单。所提出的语义深度网络也可以很容易地推广到这些任务，尽管技术细节需要进一步研究。例如，对于方面术语情感分析，语义深度网络的输入可以构造
为“[CLS]+text1+[SEP]+aspect1+[SEP]+text2+[SEP]+aspect2+[SEP]”。对于文档级情感分析，由于现有的预训练语言模型通常仅限于最多 512 个字符的序列，因此需要扩展语义深度网络的输入以处理整个文档。最后值得注意的是，开源的GML平台支持多标签因子图的构建及其逐步推理。因此，所提出的方法可以潜在地扩展到处理其他二进制甚至多标签文本分类任务。

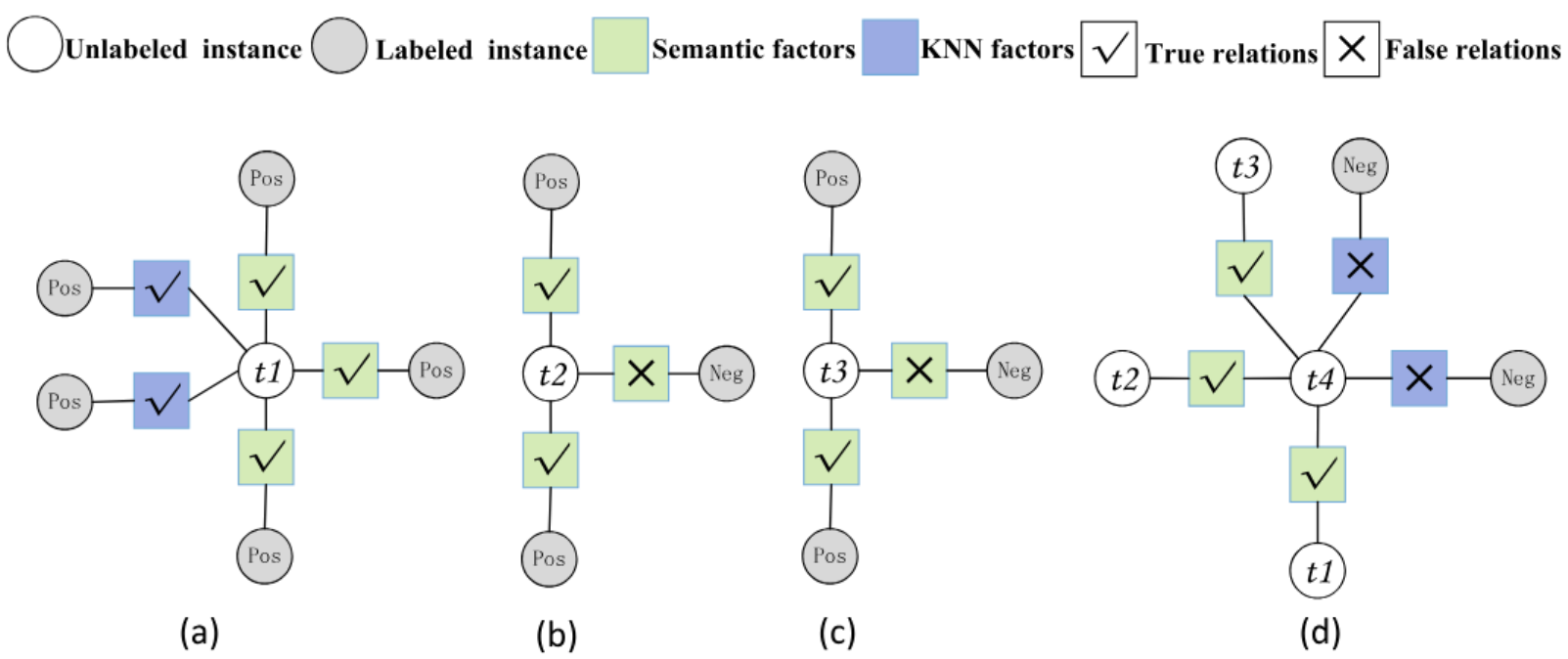

图7.渐进推理的说明性示例：1）四个子图分别表示表5中四个示例的构造因子子图； 2)真关系因子(或假关系因子)意味着其对应的极性关系为真(或假)。

k k	CR		MR		SST		推特2013	
	Acc	宏-F1	Acc	宏-F1	Acc	宏-F1	Acc	宏-F1
3 3	95.09%±0.13	96.13%±0.21	92.88%±0.20	92.76%±0.11	96.21%±0.21	96.24%±0.19	95.03%±0.47	96.56%±0.29
3 5	95.76%±0.23	96.64%±0.27	93.35%±0.17	93.22%±0.17	96.32%±0.20	96.36%±0.22	94.54%±0.43	96.21%±0.27
3 7	95.62%±0.10	96.54%±0.42	93.34%±0.35	93.22%±0.44	96.27%±0.25	96.30%±0.17	94.74%±0.35	96.35%±0.29
3 8	95.49%±0.51	96.43%±0.65	93.35%±0.48	93.22%±0.52	96.05%±0.32	96.09%±0.61	94.44%±0.33	96.14%±0.66
5 3	95.23%±0.34	96.25%±0.50	93.53%±0.32	93.46%±0.19	96.43%±0.81	96.45%±0.21	94.99%±0.34	96.53%±0.65
5 5	95.36%±0.78	96.32%±0.34	93.16%±0.56	93.03%±0.41	96.00%±0.71	96.04%±0.80	94.59%±0.91	96.23%±0.87
5 7	95.62%±0.40	96.53%±0.61	93.25%±0.21	93.14%±0.53	96.21%±0.10	96.25%±0.23	94.94%±0.31	96.48%±0.40
7 3	95.36%±0.81	96.36%±0.21	93.63%±0.30	93.56%±0.22	96.21%±0.57	96.24%±0.45	94.89%±0.18	96.46%±0.15
7 5	95.62%±0.20	96.54%±0.30	93.44%±0.82	93.36%±0.44	95.94%±0.20	95.96%±0.19	94.49%±0.21	96.17%±0.41
7 7	95.76%±0.29	96.64%±0.30	93.44%±0.29	95.94%±0.91	95.97%±0.30	96.00%±0.10	94.59%±0.90	96.24%±0.12
7 8	95.49%±0.12	96.42%±0.20	93.16%±0.51	93.03%±0.11	95.99%±0.29	96.04%±0.21	94.44%±0.34	96.14%±0.37

表 6 参数敏感性评估结果：GML 的性能敏感性 w.r.t 提取的语义关系的数量 k 以及提取的KNN关系的数量 k.

结论

在本文中，我们针对句子级情感分析任务提出了一种基于 GML 的新颖解决方案。所提出的解决方案利用现有的 DNN 模型来提取极性感知的二元关系特征，然后用于实现有效的渐进知识传递。我们对基准数据集进行的广泛实验表明，它实现了最先进的性能。我们的工作清楚地表明，渐进式机器学习与 DNN 配合进行特征提取，在句子级情感分析方面可以比纯深度学习解决方案表现得更好。

未来的工作可以从两个方向进行。首先，在许多实际场景中，准确标记的训练数据可能不容易获得。因此，在仅提供少量标记样本的弱监督环境中研究渐进式机器学习非常重要。其次，将所提出的方法扩展到其他二元甚至多标签分类任务也很有趣。

数据可用性

本研究中使用的所有数据集都是公开的。 MR 数据集可从 https://www.cs.cornell.edu/people/pabo/movie-review-data/ 获取。 CR 数据集可从 https://www.cs.uic.edu/~liub/FBS/sentiment-analy sis.html#datasets 获取。 Twitter2013 数据集可从 https://www.dropbox.com/s/byzr8yoda6bua1b/2017_English_final.zip?file_subpath=%2F2017_English_final%2FGOLD%2FSubtask_A 获取。 SST 数据集可从 http://nlp.stanford.edu/sentiment 获取。我们论文中使用的情感词典 EPA 可从 http://www.EPA 获取。indiana.edu/~socpsy/public_files/EnglishWords_EPAs.xlsx，另一个情感词典 VAD 可从 https://saifmohammad.com/WebPages/nrc-vad.html 获得。

代码可用性

我们的代码位于：https://github.com/sujingxd/SLSA-GML。

收稿日期：2023 年 4 月 21 日；接受日期：2023 年 8 月 28 日
Published online: 04 September 2023

参考

1. Bongirwar, V.K. 句子级情感分析调查。国际。 J. 计算机。科学。趋势技术。 20, 110–113 (2015)。
2. Pang, B. & Lee, L. 意见挖掘和情绪分析。成立。趋势信息。检索 20, 1–135 (2008)。
3. Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. Bert 用于语言理解的深度双向转换器的预训练 (2018)。
4. 刘，Y.等人。 Roberta 是一种稳健优化的 bert 预训练方法。 arXiv：1907.11692（arXiv 预印本）（2019）。
5. 杨，Z.等人。 Xlnet 用于语言理解的广义自回归预训练（2019）。
6. 王，Y.等人。基于渐进式机器学习的方面级情感分析。基于知识的系统212, 106509 (2021)。
7. 艾哈迈德，M.等人。用于方面术语情感分析的 Dnn 驱动渐进机器学习 (2021)。
8. Hou, B., Chen, Q., Wang, Y., Nafa, Y. & Li, Z. 用于实体解析的渐进式机器学习。 IEEE 传输。知道。数据工程34, 1803–1814 (2022)。
9. Ito, T.、Tsbouchi, K.、Sakaji, H.、Yamashita, T. 和 Izumi, K. 用于文档情感分析的上下文情感神经网络。数据科学。工程师。 5、25（2020）。
10. Mao，Y.，Zhang，Y.，Jiao，L.&Zhang，H。使用基于注意力的双向长短期记忆网络和二维卷积神经网络进行文档级情感分析。电子学 11, 25 (2022)。
11. Ranoui, M.、Mikram, M.、Yousfi, S. 和 Barzali, S. 用于文档级情感分析的 cnn-bilstm 模型。马赫。学习。知道。提炼。 1, 832–847 (2019)。
12. Rao, G.、Huang, W.、Feng, Z. 和 Cong, Q. Lstm，具有用于文档级情感分类的句子表示。神经计算 308, 49–57 (2018)。
13. Garg, S. 和 Ramakrishnan, G. Bae 基于 bert 的文本分类对抗示例 (2020)。
14. Zeng, B.、Yang, H.、Xu, R.、Zhou, W. 和 Han, X. Lcf 用于基于方面的情感分类的局部上下文聚焦机制。应用。科学。 9、3389（2019）。
15. 戴 J.、严 H.、孙 T.、刘 P. 和邱 X. 语法重要吗？ roberta (2021) 为基于方面的情感分析提供了强有力的基线。
16. Bai, X., Liu, P. 和Zhang, Y. 使用图注意神经网络研究目标情感分类的类型化句法依赖性。 IEEE 传输。音频语音语言。过程。 29, 503–514 (2021)。
17. Tripathy, A.、Agrawal, A. 和 Rath, S.K. 使用 n-gram 机器学习方法对情感评论进行分类。专家系统。应用。 57, 117–126 (2016)。
18. Fang, J. & Chen, B. 将词典知识纳入 SVM 学习以改进情感分类 (2011)。
19. Kumari, U.、Sharma, A. 和 Soni, D. 使用 SVM 分类技术对智能手机产品评论进行情感分析 (2017)。
20. Chikersal, P.、Poria, S. 和 Cambria, E. Sentu 通过将基于规则的分类器与监督学习相结合来进行推文情感分析 (2015)。
21. Wang, J., Yu, L.-C., Lai, K. R. & Zhang, X. 用于维度情感分析的树结构区域 cnn-lstm 模型。 ACM 翻译。音频语音语言。过程。 28, 581–591 (2019)。
22. Li，W.，Zhu，L.，Shi，Y.，Guo，K. 和 Cambria，E。使用词典集成两通道 cnn-lstm 系列模型进行用户评论情感分析。应用。软计算。 94, 106435 (2020)。
23. Minaee, S.、Azimi, E. 和 Abdolrashidi, A. 使用 cnn 和 bi-lstm 模型集成进行深度情感分析。 arXiv：1904.04206（arXiv 预印本）（2019）。
24. Zhou, X., Wan, X. & Shaw, J. 基于注意力的 lstm 网络用于跨语言情感分类 (2016)。
25. Li, Z.、Wei, Y.、Zhang, Y. & Yang, Q. 用于跨域情感分类的分层注意力转移网络（2018）。
26. 斯塔彭，L.等人。使用分层注意网络进行上下文建模，用于口语叙事中的情绪和自我评估情绪检测（2019）。
27. Ke, P., Ji, H., Liu, S., Zhu, X. & Huang, M. 使用语言知识进行情感感知语言表示学习（2020）。
28. Wang, S.、Fang, H.、Khabsa, M.、Mao, H. 和 Ma, H. 作为少样本学习者的蕴涵。 arXiv:2104.14690（arXiv 预印本）（2021）。
29. 曾，J.等人。使用语法感知编码器改进评论情绪分析（2019）。
30. Cheng, K., Yue, Y. & Song, Z. 基于词性和自注意力机制的情感分类。 IEEE 访问 8, 16387–16396 (2020)。

31. Reimers, N. & Gurevych, I. 使用暹罗 bert 网络的 Sentence-bert 句子嵌入 (2019)。
32. Yin, D.、Meng, T. 和 Chang, K.-W. Sentibert 是一种基于可转移变压器的组合情感语义架构（2020）。

33. 向，R.等人。神经情感分析中的情感意识。基于知识的系统226, 107137 (2021)。
34. 张，S.，Loweimi，E.，贝尔，P. 和 Renals，S。关于自注意力对于变压器自动语音识别的有用性（2021）。

35. Li, X., Li, Z., Xie, H. & Li, Q.通过自适应门合并统计特征以改进文本分类（2021）。
36. Chen, Q.、Zhang, R.、Zheng, Y. & Mao, Y. 通过标签感知数据增强进行双重对比学习文本分类。 CoRR（2022）。

致谢

该工作得到了国家自然科学基金项目（批准号：62172335、61972317、61732014、61672432）和中央高校基本科研业务费专项资金（批准号：3102019DX1004）的资助。

作者贡献

J.S.设计并实现了算法，并准备了手稿。质量控制负责制定研究计划，修改稿件。 Y.W.帮助算法实现。 Z.L.和 W.P.帮助算法设计和实证评估。 L.Z.编辑了手稿。所有作者都审阅了手稿。

利益竞争

作者声明没有竞争利益。

附加信息

信件和材料请求应发送至 J.S.