**dEFEND: Explainable Fake News Detection**

摘要

近年来，为减轻假新闻问题，对假新闻的计算检测进行了研究，产生了一些令人鼓舞的早期结果。 然而，尽管很重要，但我们认为，这项研究的关键缺失部分是这种检测的可解释性，即为什么将某条新闻检测为伪造。 因此，在本文中，我们研究了对假新闻的可解释的检测。 我们开发了一个句子注释共同注意子网络，以利用新闻内容和用户评论共同捕获可解释的前k个值得检查的句子和用户评论，以进行假新闻检测。 我们在现实世界的数据集上进行了广泛的实验，并证明了该方法不仅在F1评分方面比7个最新的假新闻检测方法明显优于至少5.33％，而且还（同时）确定了前k位用户的评论 这就解释了新闻为什么是伪造的，在NDCG中比基线好28.2％，在Precision中比基线好30.7％。

1简介社交媒体平台为用户创建，访问和共享各种信息提供了便捷的渠道。 由于社交媒体的日益普及和便利，越来越多的人在网上寻找并及时接收新闻信息。 例如，皮尤研究中心（Pew Research Center）宣布，约有68％的美国成年人在2018年从社交媒体获取新闻，而只有49％的成年人在20121年报告在社交媒体上看到新闻。 但是，与此同时，社交媒体使用户能够接触到无数的错误信息和虚假信息，包括虚假新闻，即带有故意虚假信息的新闻报道[1，40]。 例如，据一份报告估计，到2016年末总统大选2，与假新闻“ Pizzagate”相关的推文超过100万条。

如此广泛的虚假新闻具有有害的社会影响。首先，它大大削弱了公众对政府和新闻的信任。例如，具有讽刺意味的是，在2016年美国总统大选期间，针对前20名假新闻的假新闻报道范围比前20名最受关注的真实报道要大3。其次，虚假新闻可能会改变人们对合法新闻的回应方式。一项研究表明，人们在不同年龄段和政党之间对大众媒体的信任程度已大大降低4。第三，猖“的“在线”假新闻可能导致“离线”社会事件。例如，有假新闻声称巴拉克·奥巴马（Barack Obama）在爆炸中受伤，从而烧掉了1300亿美元的股票价值5。因此，能够减少假新闻在社交媒体上的传播，增强对整个新闻生态系统的信任已变得至关重要。然而，在社交媒体上检测假新闻提出了独特的挑战。首先，由于虚假新闻是故意写给误导读者的，因此仅根据新闻的内容来检测虚假新闻就并非难事。其次，社交媒体数据是大规模的，多模式的，主要由用户生成的，有时是匿名的且嘈杂的。为了应对这些挑战，最近的研究进展使用户在新闻文章上的社交参与度得以提高，以帮助推断哪些文章是假的[13，37]，并给出了一些令人鼓舞的早期结果。例如，Natali等。 [37]提出了一种混合深度学习框架，用于同时对新闻文本，用户响应和帖子来源进行建模，以进行假新闻检测。郭等。 [13]利用分层神经网络检测虚假新闻，通过选择重要用户评论的社会关注来建模用户参与。

尽管现有的基于深度学习的伪造新闻检测方法取得了成功，但是这些方法中的大多数集中于有效地检测具有潜在特征的伪造新闻，但无法解释“为什么”一条新闻被检测为伪造新闻。能够解释为什么新闻被确定为伪造是非常可取的，因为：（1）得出的解释可以提供原本对从业人员隐藏的新见解和知识； （2）从嘈杂的辅助信息中提取可解释的特征，可以进一步帮助改善假新闻检测性能。但是，据我们所知，在进行社交媒体适当解释之前，还没有尝试过计算检测假新闻的尝试。特别是，我们建议从新闻内容和用户评论的角度得出解释（参见图1）。首先，新闻内容可能包含可证实为虚假的信息。例如，记者在事实检查网站（例如PolitiFact6）上手动检查新闻报道中的声明，这通常是费时费力的工作。研究人员还尝试使用外部资源对新闻文章中的声明进行事实检查，以决定和解释新闻片段是否为伪造[6]，这可能无法检查新出现的事件（尚未经过事实检查）。其次，用户评论在社交媒体上从人群中获得了丰富的信息，包括观点，立场和观点，对检测虚假新闻很有用。例如，研究人员建议利用社交功能选择重要的评论来预测假新闻[13]。此外，新闻内容和用户评论本质上是相互关联的，并且可以提供重要的线索来解释给定新闻文章为何是假的。例如，在图1中，我们可以看到用户在评论中讨论新闻的不同方面，例如“ St.尼古拉斯是白人吗？真的是“哈哈”，它直接回应了新闻内容中的说法：“圣书总是说圣诞老人是白人。”因此，在本文中，我们通过从新闻内容和用户评论中共同探索可解释的信息来研究假新闻检测问题。为此，我们通过一个连贯的过程构建了一个可解释的假新闻检测框架，该框架包括：（1）对新闻内容进行编码的组件（通过层次化注意力神经网络来学习新闻句子表示，以捕获语义和句法线索） ；（2）对用户评论进行编码的组件（以通过单词级关注子网络学习用户评论的潜在表示形式），以及（3）句子-评论共同注意组件（以捕获新闻内容之间的相关性）和评论，并选择前k个可解释的句子和评论）。

从本质上讲，本文解决了以下挑战：（**1）如何进行可解释的假新闻检测，以同时提高检测性能和可解释性； （2）在训练过程中如何在没有事实依据的情况下提取可解释的评论； （3）如何共同建立新闻内容和用户评论之间的相关性模型，以进行可解释的假新闻检测**？ 我们针对这些挑战的解决方案产生了一个名为dEFEND（可解释的FakE新闻检测）的新颖框架。 我们的主要贡献概述如下：**•我们研究了社交媒体上可解释的虚假新闻检测的一个新问题。 •我们提供了一种原则性的方式来联合利用新闻内容和用户评论来捕获可解释的用户评论，以进行假新闻检测； •我们对现实世界的数据集进行了广泛的实验，以证明dEFEND在检测假新闻和解释假新闻结果方面的有效性**。

2相关工作

在本节中，我们简要回顾有关假新闻检测和可解释的机器学习的相关工作。

2.1假新闻检测假新闻检测方法通常着重于使用新闻内容和社交环境[40、51、52]。新闻内容特征主要从文本和视觉方面提取。文字特征捕获了假新闻内容中常见的特定写作风格[34]和耸人听闻的情感[12]。另外，使用张量因子分解[15]，深度神经网络[20、21、44]对潜在的文本表示进行建模，这可以很好地检测带有新闻内容的假新闻。从视觉元素（例如图像和视频）中提取视觉特征，以捕获假新闻的不同特征[19]。对于基于社交上下文的方法，功能主要包括基于用户，基于帖子和基于网络。从用户配置文件中提取基于用户的特征以测量其特征[3，42]。基于帖子的功能代表用户的立场[43]，主题[13]或信誉[18]。通过构建特定的网络（例如扩散网络[46]，交互网络[41]和传播网络[30、39]）可以提取基于网络的特征。最近，研究还集中在假新闻检测的挑战性问题上，例如通过对抗学习[45]和用户响应生成[35]，半监督检测[11]和无监督检测[15、49]进行假新闻早期检测。通过元属性可解释地检测假新闻[48]。在本文中，我们研究了可解释的假新闻检测的新问题，该问题旨在提高假新闻的检测性能，并同时突出显示可解释的用户评论和值得检查的新闻句子。

2.2可解释的机器学习我们的工作也与可解释的机器学习有关，通常可分为两类：**内在的可解释性和事后可解释性[**8]。 通过构建将解释性直接纳入其结构的自我解释模型，可以实现内在的解释性。 **通过发现具有大系数的特征来实现可解释性**，这些特征在解释预测中起关键作用[7]。 相反，事后可解释性要求**创建第二个模型来为现有模型提供解释**。 Koh等。 [24]建议通过影响函数来确定与给定预测结果最相关的训练点。 刘等。 提出通过归类法结构来解释网络嵌入表示[27]。

与传统的机器学习算法不同，深度学习模型（DNN）的学习表示通常无法由人类解释[8]。 因此，对深度神经网络（DNN）的解释主要着重于理解DNN中间层的神经元捕获的表示[9，26，28]。 刘等。 利用机器学习模型的解释来进行对抗性检测[28]。 Du等人建议通过导引特征反演对神经网络进行实例级解释[47]。 Karpathy等。 [22]使用字符级语言建模分析了RNN激活模式的可解释性。 研究[33]发现，RNN可以通过检查不同隐藏层的表示来学习上下文表示。 在本文中，我们建议利用共同注意机制来共同捕获新闻句子和用户评论的内在可解释性，并改善假新闻检测性能

3问题陈述

令A为新闻文章，由N句话{si} N i = 1组成。每个句子si = {wi 1，···，wii Mi}包含Mi词。令C = {c1，c2，...，cT}是与新闻A相关的T条评论的集合，其中每个评论cj = {w j 1，...，w j Qj}包含Qj个单词。类似于先前的研究[18，40]，我们将假新闻检测问题视为二进制分类问题，即每则新闻文章可以为真**（y = 1）或假（y = 0）**。同时，我们的目标是根据可辩证程度，从{si} N i = 1的所有句子中学习等级列表RS，从{cj} T j = 1的所有评论中学习排名列表RC RSk（RCk）表示第k个最能解释的句子（注释）。新闻内容中句子的可解释性表示其可检查性的程度，而评论的可解释性表示与新闻中的主要主张密切相关的用户相信新闻是假的还是真实的程度。正式地，我们可以将问题表示为可解释的假新闻检测：

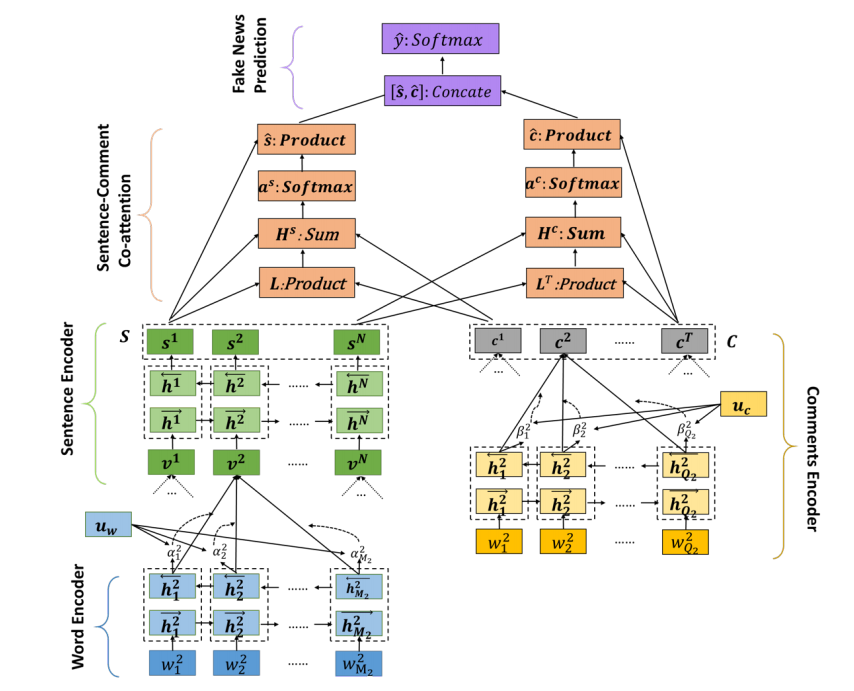
问题：可解释的伪造新闻检测。 给定一条新闻文章A和一组相关的评论C，学习一个伪造的新闻检测函数f：，以便它以可解释的句子和评论的最高预测准确性最大化 RS和RC

RS= 句子

RC=评论

4辩护：可解释的假新闻检测框架

在本节中，我们将介绍提议的可解释性假新闻检测框架的详细信息，该框架名为dEFEND（可解释的FakE新闻检测）。它由四个主要组件组成（请参见图2）：（1）新闻内容编码器（包括单词编码器和句子编码器）组件；（2）用户评论编码器组件；（3）句子注释共同注意组件； （4）假新闻预测组件。具体而言，首先，新闻内容编码器组件通过分层的单词和句子级编码来描述从新闻语言特征到潜在特征空间的建模。接下来，用户评论编码器组件说明了通过单词级关注网络提取评论潜在特征的过程；然后，句子注释共同注意组件模拟新闻句子和用户注释之间的相互影响以学习特征表示，并通过共同注意学习中的注意权重来学习句子和注释的可解释性程度。最后，假新闻预测组件显示了将新闻内容和用户评论功能进行串联以进行假新闻分类的过程。



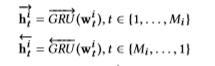
提议的dEFEND框架由四个组件组成：（1）新闻内容（包括单词级和句子级）编码器，（2）用户评论编码器，（3）句子-注释共同注意组件，以及（ 4）假新闻预测组件。

4.1新闻内容编码

由于故意制作虚假新闻来传播不准确的信息，因此它们往往具有自以为是和耸人听闻的语言风格，有可能帮助发现虚假新闻。此外，新闻文档包含不同级别的语言提示，例如单词级别和句子级别，这对于新闻为什么是伪造的可解释性提供了不同程度的重要性。例如，在虚假新闻声明“便士：米歇尔·奥巴马是我们有史以来最粗俗的第一夫人”中，“粗俗”一词提供了更多的信号来决定新闻声明是否是虚假的，而不是句子中的其他词语。

最近，研究人员发现，层次化注意力神经网络[50]对于突出显示重要词或句子进行分类的文档表示[4]来说非常实用和有用。它采用分层神经网络通过自我注意机制对词级和句子级表示进行建模。受[4]的启发，我们建议通过分层结构来学习新闻内容的表示形式。具体而言，我们首先通过注意使用单词编码器来学习句子向量，然后通过句子编码器组件来学习句子表示。

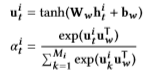
4.1.1字编码器。 我们通过基于递归神经网络（RNN）的单词编码器学习句子表示。 尽管从理论上讲，RNN能够捕获长期依赖关系，**但实际上，随着序列变得更长，旧的内存将逐渐消失。** 为了捕获RNN的长期依赖性，**使用门控循环单元（GRU）[5]来确保更持久的内存**。 类似于[50]，我们采用GRU编码单词序列。 为了进一步捕获注释的上下文信息，我们使用双向GRU [2]从单词的两个方向建模单词序列。 双向GRU包含从单词w i 1到w i Mi读取句子si的前向GRU-→f和从单词w i Mi到w i 1读取句子si的后向GRU←-f：



我们通过将前向隐藏状态-→hi t和后向隐藏状态←-hi t，即hi t = [-→hi t，←-hi t]串联起来，获得单词wit的注释，其中包含以下信息： 请注意，**并非所有单词都对句子含义的表达有同等的作用。因此，我们引入了注意力机制来学习衡量单词重要性的权重**，并如下计算句子向量vi∈R2d×1



其中αit衡量第t个单词对句子si的重要性，而αit的计算如下：



其中**αi t衡量句子si中第t个单词的重要性**，u i t是通过将隐藏状态h i t馈送到完全嵌入层而获得的h i t的隐藏表示，而**uw是表示世界级上下文向量的权重参数**。

4.1.2句子编码器。

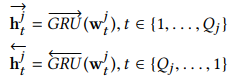
与单词编码器类似，我们利用带有GRU单元的RNN对新闻中的每个句子进行编码。 我们在句子级别捕获上下文信息，以从学习的句子向量vi中学习句子表示形式h i。 具体来说，我们可以使用双向GRU对句子进行编码，如下所示：



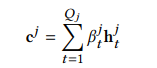
我们通过串联前后隐藏状态获得si∈R 2d×1，即s i = [-→h i，←− h i]，它从句子si周围的相邻句子中捕获上下文。

4.2用户评论编码

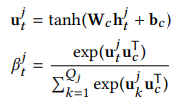
人们通过社交媒体帖子表达对伪造新闻的情感或看法，例如怀疑的意见，轰动的反应等评论。这些文本信息已被证明与原始新闻的内容有关。因此，评论可能包含有用的语义信息，有可能有助于伪造新闻检测。接下来，我们演示如何编码注释以学习潜在表示。从社交媒体提取的评论通常是短文本，因此我们使用RNN直接在评论中编码单词序列，以学习评论的潜在表示形式。与单词编码器类似，我们采用双向GRU对注释中的单词序列进行建模。具体来说，给定带有单词wjt，t∈{1，···，Qj}的注释cj，我们首先将每个单词wjt映射到单词向量wjt ∈R d具有嵌入矩阵。然后，我们可以获得前馈隐藏状态−→h j t和后向隐藏状态←− h j t，如下所示：



我们通过将-→h j t和←-h j t进行级联来进一步获得单词w j t的注释，即h j t = [-→h j t，←-h j t]。 我们还引入了注意力机制来学习衡量每个单词的重要性的权重，注释向量c j∈R 2d的计算如下：



其中βj t度量第t个单词对注释cj的重要性，并且βj t的计算如下：



其中，u j t是通过将隐藏状态h j t馈送到完全嵌入层而获得的h j t的隐藏表示，而uc是权重。

4.3句子与评论的共同注意

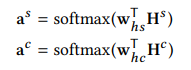
我们注意到新闻内容中并非所有句子都是假的，实际上，有些句子是正确的，但仅用于支持错误的声明句子[10]。因此，新闻句子在确定和解释某条新闻是否为假新闻方面可能并非同等重要。例如，“米歇尔·奥巴马（Michelle Obama）如此低俗，她不仅在说话而已。”这句话与虚假说法“便士：米歇尔·奥巴马（Michelle Obama）是我们有史以来最粗俗的第一夫人”紧密相关，而“第一夫人谴责” “共和党总统候选人”表达了一些事实，在发现和解释该新闻是否为假新闻方面帮助较小。同样，用户评论可能包含有关重要方面的相关信息，这些重要方面可以解释新闻为什么是伪造的，同时它们的信息量也较小，且噪音较大。例如，有一条评论“便士在哪里说呢？今天早上我在CBS上看到他，他没有说这些话。”比起“便士绝对正确”之类的其他评论，它更容易解释和发现虚假新闻。因此，我们旨在选择新闻句子和用户评论，以解释新闻为什么是假的。由于它们提供了很好的解释，因此它们也应有助于发现虚假新闻。这建议我们设计注意力机制，使新闻句子和评论的表示具有较高的权重，这对伪造新闻检测很有帮助。具体来说，我们使用句子注释大写，是因为它可以捕获句子和注释的语义相似性，并进一步帮助同时学习句子和注释的注意力权重。我们构造新闻句S=[s1,…,sn]∈R 2d×N的特征矩阵和用户评论的特征图C = {c 1，···，c T}∈R 2d×T，共同关注与句子和评论有关 同时。类似于[29]，我们首先计算亲和度矩阵F∈R T×N，如下所示：



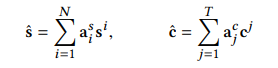
其中Wl∈R 2d×2d是要通过网络学习的权重矩阵。 遵循[29]中的优化策略，我们可以将亲和度矩阵视为一个特征，并学会预测句子和评论注意图，如下所示：



其中Was，Wc∈R k×2d是权重参数。 句子和注释的注意力权重计算如下：



其中s∈R 1×N和c∈R 1×T分别是每个句子s i和注释c j的注意力概率。 whs，whc∈R 1×k是权重参数。 相似度矩阵F将用户评论关注空间转换为新闻句子关注空间，反之亦然。 根据以上注意力权重，将评论和句子的注意力向量计算为评论特征和句子特征的加权总和，即



其中sˆ∈R 1×2d和cˆ∈R 1×2d是新闻句子和用户注意通过共同注意学习的特征。

4.4建议的框架：dEFEND

我们介绍了如何通过对单词级别和句子级别的层次结构进行建模来对新闻内容进行编码，如何通过单词级别的关注网络对评论进行编码以及如何对共同注意进行建模以学习句子和句子。 评论表示。 我们将这些组件进一步整合在一起，并根据以下目标预测假新闻：



其中yˆ = [yˆ0，yˆ1]是预测概率向量，其中yˆ0和yˆ1分别表示标签的预测概率为0（真实新闻）和1（虚假新闻）。 y∈{0，1}表示新闻的地面真相标签。 [sˆ，cˆ]表示新闻句和用户评论的学习功能的串联。 bf∈R 1×2是偏差项。 因此，对于每个新闻，目标是如下最小化交叉熵损失函数，



其中θ表示网络参数。

网络中的参数通过RMSprop进行学习，**RMSprop是一种自适应学习速率方法**，该方法将学习速率除以平方梯度的指数衰减平均值。 我们选择RMSprop作为优化器，因为它是一种异常流行的有效确定学习速率的方法，该方法**广泛用于训练神经网络**。

5实验

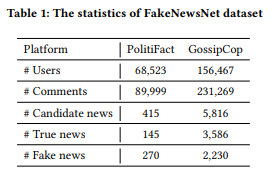
在本节中，我们介绍实验以评估提出的dEFEND框架的有效性。 具体来说，我们旨在回答以下评估问题：

**•EQ1 dEFEND是否可以通过同时对新闻内容和用户评论进行建模来改善假新闻分类性能？ •EQ2新闻内容和用户评论分别在提高dEFEND的检测性能方面有多有效？**

**•EQ3可以dEFEND捕获新闻句子和用户评论，这些新闻句子和用户评论可以解释为什么一条新闻是假的**

5.1数据集

我们利用一种称为FakeNewsNet的全面的假新闻检测基准数据集[38，40]。该数据集是从两个经过事实检查的平台收集的：GossipCop和PolitiFact，都包含带有标签的新闻内容和社交环境信息。新闻内容包括新闻的元属性（例如，正文），社交上下文包括新闻项目的相关用户社交活动（例如，Twitter中的用户评论）。请注意，我们保留至少3条评论的新闻片段。表1中显示了数据集的详细统计信息。



5.2比较伪造新闻检测方法

代表性的最新伪造新闻检测算法如下：

•RST [36]：RST代表修辞结构理论，它表示：建立一个树形结构来表示文本中单词之间的修辞关系。 RST可以通过将修辞关系的频率映射到向量空间7来提取新闻样式特征。

•LIWC [32]：LIWC代表语言查询和字数统计，广泛用于提取属于心理语言学类别的词典。它从心理学和欺骗的角度学习特征向量8。

•HAN [50]：HAN在新闻内容上利用层次化注意力神经网络框架进行伪造新闻检测。它对新闻内容进行编码，使每个句子的单词级别注意和每个文档的句子级别注意成为可能。

•text-CNN [23]：text-CNN利用卷积神经网络对新闻内容进行建模，可使用多个卷积过滤器捕获不同粒度的文本特征。

•TCNN-URG [35]：TCNN-URG由两个主要组件组成：一个两级卷积神经网络，用于从新闻内容中学习表示形式；以及一个条件变分自动编码器，用于从用户评论中捕获特征。

•HPA-BLSTM [13]：HPA-BLSTM是一种神经网络模型，可以通过社交媒体上用户参与的字级，后级和子事件级的层次注意力网络学习新闻表示。此外，还提取了帖子特征，以学习帖子级别的注意力权重。

•CSI [37]； CSI是一种混合式深度学习模型，它利用来自文本，响应和源的信息。 新闻表示是通过LSTM神经网络建模的，其中Doc2Vec [25]嵌入新闻内容和用户评论作为输入，并且为了公平比较，忽略了用户功能。

请注意，为了进行公平的比较，我们选择了上述伪造的新闻方法，这些方法从以下几个方面提取了特征：

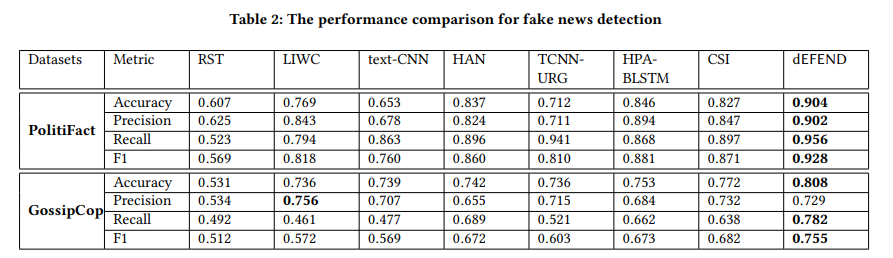
（1）仅新闻内容，例如RST，LIWC，text-CNN，HAN；

（2）仅用户评论，例如HPA-BLSTM，以及

（3）新闻内容和用户评论，例如TCNN-URG和CSI。

对于诸如RST和LIWC之类的特征提取方法，我们将它们输入到不同的学习算法中，然后选择能够获得最佳性能的算法。 该算法包括逻辑回归，朴素贝叶斯，决策，决策树和随机森林。 我们使用带有默认参数设置的scikit-learn [31]运行这些算法。

5.3假新闻检测性能



为了回答EQ1（**dEFEND是否可以通过同时对新闻内容和用户评论进行建模来改善假新闻分类性能？**），我们首先将dEFEND与5.2节中介绍的代表性假新闻检测算法进行比较。要评估伪造新闻检测算法的性能，我们使用以下指标，这些指标通常用于评估相关领域中的分类器：准确性，准确性，召回率和F1。我们随机选择75％的新闻文章进行培训，其余25％的新闻进行测试，该过程执行了5次，平均效果记录在表2中。从该表中，我们观察到以下几点：

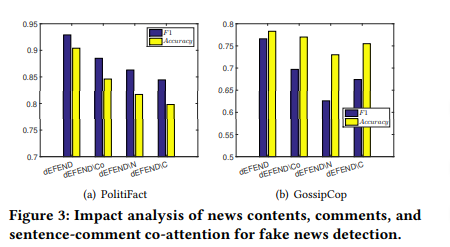
•对于新闻内容，方法RST，LIWC和HAN，我们看到两个数据集的HAN> LIWC> RST。这表明：1）HAN可以通过新闻内容中的层次注意力神经网络更好地捕捉句法和语义线索，以区分假新闻和真实新闻。 2）LIWC可以更好地捕捉新闻内容中的语言特征。 LIWC的良好结果表明，伪新闻与真实新闻在选择显示心理计量学特征的词语方面有所不同。

•此外，同时使用新闻内容和用户评论的方法比仅基于新闻内容的方法和仅基于用户评论的方法（即dEFEND> HAN或HPA-BLSTM和CSI> HAN或HPA-BLSTM）要好。这表明从新闻内容中提取的特征和相应的用户评论具有互补信息，从而提高了检测性能。

•此外，基于用户评论的方法的性能比基于新闻内容的方法稍好。例如，在PolitiFact和Gossipcop数据的准确性和F1方面，我们具有HPA − BLSTM> HAN。它表明，从用户评论中提取的功能比仅用于预测假新闻的新闻内容具有更大的区分能力。

•通常，对于基于新闻内容和用户评论的方法（即dEFEND，CSI和TCN N -U RG），我们可以看到dEFEND始终优于CSI和TCN N-U RG，即dEFEND> CSI> TCN N – U RG，就两个数据集的所有评估指标而言。例如，与CSI的准确度和F 1得分相比，dEFEND在PolitiFact上的平均相对改善为4.5％，在Gossipcop上的平均相对改善为4.7％，在10.8％。它支持对新闻句子和用户评论进行共同注意建模以进行假新闻检测的重要性

5.4评估新闻内容和用户评论的影响



除新闻内容外，我们还从用户评论中捕获信息，并将其与新闻内容集成在一起，并引起关注。为了回答EQ2(**新闻内容和用户评论分别在提高dEFEND的检测性能方面有多有效？**)，我们通过定义dEFEND的三个变体来进一步研究这些组件的影响：

•dEFEND \ C：dEFEND \ C是dEFEND的变体，无需考虑用户评论中的信息。它首先对新闻内容进行编码，使每个句子的单词级别受到关注，然后通过平均池化层对得到的句子特征进行平均，然后馈入softmax层进行分类。

•dEFEND \ N：dEFEND \ N是dEFEND的变体，不考虑新闻内容中的信息。它首先利用注释编码器学习注释特征，然后将结果注释特征通过平均池化层平均，然后馈入softmax层进行分类。

•dEFEND \ Co：dEFEND \ Co是dEFEND的变体，它消除了句子注释共同注意。取而代之的是，它对句子和评论分别进行自我关注，并将结果特征连接到一个密集层，并馈入softmax层进行分类。所有变体中的参数都是通过交叉验证确定的，并且在图3中报告了最佳性能。我们观察到以下几点：

•当我们消除对新闻内容和用户评论的共同关注时，性能会降低。它暗示了对相关性进行建模的重要性，并捕获了新闻内容和用户评论之间的相互影响。

•当我们消除新闻内容的影响时，与dEFEND相比，dEFEND \ N的性能会下降。例如，就PolitiFact而言，性能在F1和准确性指标方面分别降低了4.2％和6.6％，在GossipCop中降低了18.2％和6.8％。结果表明，dEFEND中的新闻内容很重要。

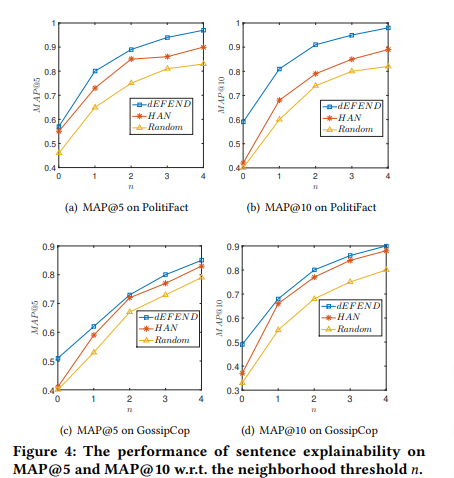
•当消除用户评论的影响时，对于dEFEND \ C有类似的观察结果。结果表明，考虑用户评论的功能以指导dEFEND中假新闻检测的重要性。通过对dEFEND的成分分析，我们得出以下结论：（1）新闻内容和用户评论的成分都可以促进dEFEND假新闻检测性能的提高； （2）有必要对新闻内容和用户评论进行建模，因为它们包含补充信息。

5.5可解释性评估和案例研究

在本小节中，为了回答EQ 3(**可以dEFEND捕获新闻句子和用户评论，这些新闻句子和用户评论可以解释为什么一条新闻是假的**)，我们从新闻句子和用户评论的角度评估dEFEND框架的可解释性性能。值得一提的是，5.2中的所有基线方法都是为伪造新闻检测而设计的，最初都没有提议发现可解释的新闻句子或用户评论。为了测量dEFEND的可解释性性能，我们选择HAN来比较新闻句子的可解释性，并选择HPA-BLSTM作为用户评论可解释性的基准，因为他们可以分别学习新闻句子和用户评论的关注权重。请注意，HAN使用注意力机制来学习文档结构，而HPA-BLSTM使用注意力机制来学习注释的时间结构。由于文档中没有时间结构，因此不能在注释中使用HAN。同样，文档结构中也没有时间关系，因此HPA-BLSTM无法直接应用于新闻内容。

5.5.1新闻句子的可解释性。

在本小节中，我们演示新闻句（RS）的可解释性等级列表的性能。具体来说，我们想了解由我们的方法确定的排名最高的可解释句子是否更可能与假新闻中值得检查的主要主张相关，即值得检查。因此，我们利用ClaimBuster [14]来获得一条新闻内容中所有可检查的句子的基本事实等级列表RS〜。 ClaimBuster提出了一种评分模型，该模型利用了各种语言学特征，这些特征是使用过去的大选辩论中成千上万的句子（由人类编码员标记）进行训练的，并给出了0到1之间的“可检查性”得分。得分越高，越可能该句子包含值得检查的事实主张。分数越低，句子越不真实，主观和自以为是。我们使用评估指标MAP将dEFEND（RS（1））和HAN（RS（2））的新闻内容中可解释句子的前k个排名列表与ClaimBuster的前k个排名列表RS〜进行比较。 @k（平均平均精度），其中k设置为5和10。我们还引入了另一个参数n，该参数控制窗口大小，当将RS（1）和RS（2）中的句子与RS〜中的前k个句子中的每一个。



从图4，我们观察到以下几点：

•通常，对于在两个数据集的新闻内容中查找值得检查的句子，我们可以看到dEFEND> HAN> Random。这表明dEFEND中的句子注释共同注意组件可以帮助选择更多值得检查的句子。 •随着n的增加，我们放宽了条件以匹配基本事实中的可检查句子，因此MAP性能不断提高。

•当n = 1时，对于PolitiFact，dEFEND在MAP @ 5和MAP @ 10上的性能提高到超过0.8，这表明dEFEND可以很好地在RS〜中地面真理句子的1个相邻句子中检测到可检查的句子。 5.5.2用户注释的可解释性。

我们使用Amazon Mechanical Turk（AMT）9部署了多个任务，以评估假新闻的评论RC的可解释性等级列表。我们执行以下设置来为总共50个虚假新闻片段部署AMT任务。对于每篇新闻文章，我们首先过滤掉少于50个单词的简短文章。此外，对于内容超过500个单词的超长文章，我们仅显示前500个单词，以减少工作人员的阅读量。由于新闻文章的前3-4段通常会概述内容，因此前500个字通常足以捕捉新闻的要旨。然后，我们招募了位于美国的AMT工人（他们更可能熟悉文章的主题），其批准率> 0.95。为了评估用户评论的可解释性，对于每则新闻，我们有两个前k个评论列表，L（1）=（L（1）1，L（1）2，··，L（1）k ），将dEFEND用于HPA-BLSTM，而L（2）=（L（2）1，L（2）2，...，L（2）k）。使用从高到低的注意力权重，选择并排序顶部的注释。为了评估选择最高可解释注释的模型能力，我们根据经验设置k =5。我们部署了两个AMT任务来评估可解释排名性能。

对于任务1，我们执行按列表比较。我们要求工人在L（1）和L（2）之间选择总体上更好的清单。为了消除位置偏差，我们在呈现给工人时随机分配L（1）和L（2）的位置（顶部和底部）。我们让每个工作人员为每个新闻片段在L（1）和L（2）之间选择更好的列表。我们确保每个新闻都由3名工人进行评估，并最终获得150个工人选择的结果。在工人级别，我们计算选择L（1）和L（2）的工人数量，并计算他们的获胜率（简称WR）。在新闻级别，我们为每个新闻对所有3名工人进行多数表决，并确定工人选择L（1）还是L（2）。对于每个新闻，我们还通过计算L（1）与L（2）之间的比率来计算工作人员级别的选择。从图5中，我们得出以下观察结果：

•在工作人员级别和新闻级别，dEFEND可以选择比HPABLSTM更好的top-k可解释注释。首先，在工人级别中，在150名工人中（WR = 0.65），有98名工人选择L（1）胜过L（2）。其次，在新闻级别，dEFEND在50个新闻片段中有32个（WR = 0.64）比HPA-BLSTM更好。

•我们可以看到，有更多的新闻片段，使3名工人对L（1）（3 vs 0）的投票比起相反的情况（0 vs 3）一致，即14> 7。同样，在更多情况下，有2名工人投票赞成dEFEND，而不是HPA-BLSTM，即18> 11。

对于任务2，我们执行逐项评估。对于L（1）和L（2）中的每个注释，我们要求工作人员从{0，1，2，3，4}中选择一个分数，其中0表示“根本无法解释”，1表示“无法解释， ” 3表示“有些可解释”，4表示“高度可解释”，2表示“介于两者之间的某处”。为了避免由于不同的用户标准引起的偏见，我们将L（1）和L（2）中的评论顺序进行了打乱，并要求工作人员评估每个评论对新闻的可解释性。为了估计评论的感知等级的可解释性（即，具有较高等级的可解释评论比较低等级的评论更可取），我们使用NDCG（归一化累积增益）[16]和Precision @ k作为评估指标。 NDCG广泛用于信息检索中，以测量搜索引擎中的文档排名性能。它可以通过将建议的排名与通过用户反馈测得的理想排名列表进行比较来衡量排名。 Precision @ k是在前k个集合中相关的推荐项的比例。同样，我们确保每个新闻都由3名工作人员评估，并且每种方法总共获得750个工作人员评级的结果。结果显示在图6中，其中新闻文章按两种方法之间的指标差异按降序排序（例如NDCG（dEFEND）-NDCG（HPA-BLSTM））。我们只显示Precision @ 5的结果，因为Precision @ 10的结果相似。我们有以下观察结果：

•在50篇虚假新闻文章中，在逐项评估方面，dEFEND在38例案例中获得的NDCG得分高于HPA-BLSRM。 dEFEND和HPA-BLSRM在50例以上的NDCG总平均分值分别为0.71和0.55。

•在Precision @ 5上可以找到类似的结果。 dEFEND在35条虚假新闻方面优于HPA-BLSTM，在7条上并列。 dEFEND和HPA-BLSRM在50例中的Precision @ 5总体平均得分分别为0.67和0.51。

案例分析。

我们将dEFEND与HPA-BLSTM进行了比较，并展示了我们正确地排名较高但被HPA-BLSTM忽略的可解释注释，如图7所示。我们可以看到：（1）dEFEND可以将比不可解释注释更高的可解释注释排名。例如，评论“ ...总统无权授予公民身份...”被排在首位，这可以准确地解释为什么为什么“总统授予2500名伊朗人包括政府官员的家属的美国公民身份”一词的原因。新闻内容是假的； （2）与那些干扰和无关的评论相比，我们可以赋予可解释的评论更高的权重，这可以帮助选择更多的相关评论来帮助检测假新闻。例如，无关的评论“从他们的身边走开...”的关注权重为0.0080，小于解释权重为0.0086的可解释的评论“不是嫁接和回报通常为犯罪”，因此选择后一个评论以成为假新闻预测的更重要功能。

6结论和未来工作

假新闻检测近年来受到越来越多的关注。但是，了解为什么一条新闻被检测为伪造也很重要。我们研究了可解释的假新闻检测的新问题，其目的是：

1）显着提高检测性能；

2）发现可解释的新闻句子和用户评论，以了解为什么新闻被识别为伪造。

我们提出了一个深度分层的共同注意网络，以学习用于假新闻检测和可解释的句子/注释发现的特征表示。在现实世界数据集上的实验证明了所提出框架的有效性。

为了以后的工作，首先，我们可以结合新闻记者专家或事实检查网站的事实检查内容，以进一步指导学习过程，以获得值得检查的新闻句子。其次，我们将探索如何使用其他用户参与作为辅助信息（例如喜欢）来帮助发现可解释的评论。第三，我们可以考虑发布可解释评论的用户的信誉，以进一步提高假新闻检测性能。

8关于可重复性的附录

在本节中，我们提供实验设置和配置的更多详细信息，以使我们的工作具有可重复性。

8.1假新闻检测

我们将建议的框架dEFEND与5.2节中讨论的7种基线方法进行了比较，包括RST，LIWC，text-CNN，HAN，TCNN-URG，HPA-BLSTM和CSI。 通过以下链接，可在“ Fake new detection”文件夹下找到我们已实现的所有代码：https：//tinyurl.com/ybl6gqrm。 其他代码如下：

• RST: we used the publicly available implementation for paper [17]: <https://github.com/jiyfeng/DPLP>

• LIWC: we used the publicly available tool at: http://liwc. wpengine.com/

• text-CNN: we used the publicly available implementation at: <https://github.com/dennybritz/cnn-text-classification-tf>

• HAN: we used the publicly available implentation at: https: //github.com/richliao/textClassifier

• TCNN-URG: we implemented this algorithm based on the description in the paper [35], and shared the code, named as tcnn.py, in the above link

• HPA-BLSTM: we used the implementation provided by the authors of [13]

• CSI: we used the implementation available at: https://github. com/sungyongs/CSI-Code

• dEFEND: 我们在Python中实现了算法-主算法为defend.py，数据处理为go\_defend.py-并在上述链接中共享了它们。

对于数据集，我们还使用了公共可用的数据集FakeNewsNet [38]，该数据集可从以下网址获得：https：//github.com/KaiDMML/FakeNewsNet。 对于dEFEND的参数设置，我们介绍了主要参数设置的详细信息，如表3所示。主要参数的说明如下：

•MAX\_SENTENCE\_LENGTH：控制新闻句子的最大长度的阈值

•MAX\_SENTENCE\_COUNT：控制句子的最大数目的阈值

•MAX\_COMMENT\_LENGTH：控制用户评论的最大长度的阈值

•MAX\_COMMENT\_COUNT：控制用户最大数目的阈值 注释

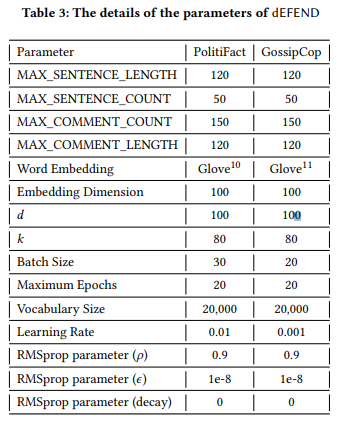
•Vocabulary Size：控制词汇最大量的阈值

•Embedding Dimension：嵌入层的维度

•Word Embedding：用于初始化词向量的词修饰包

•d：BLSTM隐藏状态的量

•k： 注意图的大小，如等式中所示。



8.2新闻和评论的可解释性

我们将详细阐述如何评估实验中句子和评论的可解释性。

8.2.1新闻句子。

我们从在线工具ClaimBuster（具有默认设置）获得了可检查句子的基本事实。 ClaimBuster的网址为：<https://idir-server2.uta.edu/claimbuster/>

8.2.2用户注释。

我们将介绍我们在Amazon Mechanical Turk部署的两个任务的详细信息。

•任务1。我们为每个假文章提供了dEFEND和HPA-LSTM识别的两个topk注释列表，并让工作人员选择可以“集体”更好地解释为什么这是假新闻的列表。 为了消除位置偏差，我们在顶部和底部之间随机调整了两个列表的顺序。 每个人类智能任务（HIT）包含五篇伪造的文章，并分配给三名不同的工作人员。 任务1的HIT屏幕截图如图8所示。为了确保众包任务的质量，我们对AMT工作人员设定了两个要求：（1）批准的工作人员分配百分比应大于95％； （2）工人的所在地应在美国

•任务2。我们分别测试了dEFEND和HPA-LSTM检测到的用户评论的可解释性。 我们为每个假新闻文章提供了由两种方法识别的“混合”评论列表，并让工作人员为每个评论分配0-4的评分，其中0表示“根本无法解释”，1表示“没有太多解释” ，“ 3”表示“可解释”，“ 4”表示“高度可解释”，“ 2”表示“介于两者之间”。 请注意，为了消除位置偏差，我们随机调整了评论的顺序，此外，为了确保任务的质量，我们应用了与任务1相同的两个要求。每个HIT都有一个假物品，并被分配了 到3个不同的工作人员，任务2的HIT屏幕截图如图9所示。