GCAN:用于社交媒体上可解释的假新闻检测的图形感知共同注意网络

Yi-Ju Lu台
湾台南市国立成功大学统计系
l852888@gmail.com

Cheng-Te Li国
立成功大学数据科学研究所 台
湾台南 chengte@mail.ncku.edu.tw

抽象的

本文解决了社交媒体上更现实场景下的假新闻检测问题。给定源短文本推文和相应的没有文本评论的转发用户序列,我们的目标是预测源推文是否是假的,并通过突出显示可疑转发者的证据和他们关注的词来生成解释。我们开发了一种新颖的基于神经网络的模型,即图形感知共同注意网络 (GCAN),以实现该目标。在真实推文数据集上进行的大量实验表明,GCAN 的准确率平均比最先进的方法高出 16%。此外,案例研究还表明 GCAN 可以产生合理的解释。

1 简介

社交媒体在人们的日常生活中不可或缺,用户可以在其中表达自己、获取新闻并相互互动。信息可以通过社交网络进一步传播。对源故事的观点和情绪可以通过用户参与和互动来反映。社交网络便捷和低成本的本质带来了集体智慧,但同时也带来了负面的副产品,即假新闻等错误信息的传播。

假新闻是一种在社交媒体上故意包含虚假信息的新闻故事（Rashkin 等人， 2017 年； Allcott 和 Gentzkow， 2017 年）。假新闻的广泛传播会误导公众,并为某些政党带来不公正的政治、经济或心理利益（Horne 和Adali， 2017 年； Allcott 和 Gentzkow， 2017 年）。数据挖掘和机器学习技术被用来检测假新闻（Shu et al., 2017; Cha et al., 2020）。典型的方法依赖于新文章的内容来提取文本特征,例如作为 n-gram 和词袋,并将监督学习（例如,随机森林和支持向量机)应用于二元分类（Shu et al., 2017）。

NLP 研究人员还学习高级语言特征,例如主动/断言动词和主观性(Popat, 2017)以及写作风格和一致性(Potthast et al., 2018)。还研究了多模式上下文信息,例如用户配置文件（Yang 等人， 2012 年； Liu 和 Wu， 2018 年)和转发传播（Ruchansky 等人， 2017 年； Shu等人， 2019a）。

尽管如此,在线检测假新闻仍然面临严峻挑战。首先,现有的基于内容的方法(Castillo et al., 2011; Potthast et al., 2018; Shu et al., 2019a)要求文档是长文本,例如新闻文章,以便单词和句子的表示可以更好地学习。然而,社交媒体上的推文通常是短文本（Yan et al., 2015）,这会产生严重的数据稀疏性问题。其次,一些最先进的模型（Ruchansky 等人， 2017 年； Liu 和Wu， 2018 年； Shu 等人， 2019a）需要为每个新闻故事收集丰富的用户评论,以了解转发者的意见,通常为识别假新闻提供有力证据。然而,社交媒体上的大多数用户倾向于简单地转发源故事而不留下任何评论（ Kwak 等人， 2010 年）。第三,一些研究(Ma et al., 2018)认为社交网络中的信息级联（即转发)路径有助于对错误信息进行分类,从而学习基于树的传播结构的表示。然而,大多数时候获取转发的扩散结构是昂贵的

由于隐私问题（Li 等人， 2018 年）。许多用户选择隐藏或删除社交记录

相互作用。第四,如果服务提供商或政府机构希望检查谁是支持假新闻的可疑用户,他们在制作假新闻时关注哪些话题

news (Reis et al., 2019),现有模型无法提供解释。尽管 dEFEND (Shu et al., 2019a)可以生成合理的解释,但它需要源文章的长文本和用户评论的文本。

本文涉及社交媒体上更现实场景下的假新闻检测。我们

预测源推文故事是否是假的,仅给出其短文本内容和用户的转发序列,以及用户配置文件。也就是说,我们在三种设置下检测假新闻:(a) 短文本源推文,(b) 没有用户评论的文本,以及 (c)没有社交网络和 diffu 的网络结构

锡安网络。此外,我们要求假新闻检测模型能够解释,即在确定故事是假的时突出证据。该模型有望指出支持假新闻传播的可疑转发者,并从源推文中突出显示他们特别关注的词。

为了实现这个目标,我们提出了一个新的模型， 1 图感知共同注意网络(GCAN) 我们首先从他们的个人资料和社交互动中提取用户特征,并从源短文本中学习词嵌入。然后我们使用卷积和递归神经网络来学习基于用户特征的转发传播的表示。构建图来模拟用户之间的潜在交互,图卷积网络用于学习用户交互的图感知表示。我们开发了一种双重共同注意机制来学习源推文和转推传播之间的相关性,以及源推文和用户交互之间的共同影响。二进制预测是基于学习到的嵌入生成的。

我们将贡献总结如下。 (1) 我们研究了一种新颖且更现实的社交媒体假新闻检测场景。 (2) 为了准确检测,我们开发了一个新模型 GCAN,以更好地学习用户交互、转发传播及其与源短文本的相关性的表示。 (3) 我们的双重共同注意机制可以产生合理的解释。

(4) 与最先进的模型相比,对真实数据集的广泛实验证明了 GCAN 的有前途的性能。案例研究中也展示了GCAN 的可解释性。

1GCAN模型的代码可用,可以ac 通过以下方式终止： https: //github.com/l852888/GCAN

我们将本文组织如下。第2节回顾了社交媒体中假新闻检测的相关方法。我们在第 3 节中描述了问题陈述。然后在第 4 节中,将详细阐述我们提出的 GCAN 模型的细节。

第5节展示了评估设置和结果。我们在第 6 节中总结了这项工作。

2 相关工作

基于内容的方法依靠文本内容来检测新闻文章的真实性,通常指的是长文本。监督学习研究了各种文本特征,包括 TF-IDF 和主题特征（Castillo等人， 2011 年）、语言风格（例如,词性、事实/断言动词和主观性） （Popat， 2017 年）)、写作风格和一致性(Potthast et al., 2018)和社会情绪(Guo et al., 2019)。

赵等。 (2015)发现用户响应中的查询短语很有用， Ma 等人。 (2016)使用递归神经网络来学习更好的用户响应表示。

基于用户的方法对转发源故事的用户的特征进行建模。杨等。 (2012)提取基于帐户的特征,例如“已验证”、性别、家乡和关注者数量。舒等。 (2019b)揭示虚假新闻和真实新闻之间的用户配置文件有很大不同。 CRNN （Liu和 Wu， 2018 年)设计了一种联合循环和卷积网络模型 (CRNN),以更好地表示转发者的资料。基于会话的异构图嵌入(Jiang et al., 2018)被提出来学习用户的特征,以便他们可以在共享帐户中被识别。然而,由于这种方法依赖于会话信息,因此不能直接应用于假新闻检测。

基于结构的方法利用社交网络中的传播结构来检测假新闻。桑普森等人。 (2016)利用隐含信息,即主题标签和 URL,连接用户没有社交链接的对话,发现此类隐含信息可以提高谣言分类的性能。马等。 (2017)创建了一种基于内核的方法,可以捕获区分不同类型谣言的高阶模式。

马等。 (2018)开发树结构递归神经网络来学习谣言传播结构的嵌入。尽管多关系图嵌入方法(Feng et al., 2019; Wang and Li, 2019)能够有效地学习不同类型的实体（与源新闻相关）

表 1:相关研究的比较。列符号:新闻故事文本 (NS)、回复评论 (RC)、用户特征 (UC)、传播结构 (PS)、社交网络 (SN) 和模型可解释性 (ME)。

对于 NS 列，“S”和“L”分别表示短文本和长文本。

	NS RC UC PS SN 我
马等。 (2016)	(S)
马等。 (2018)	(S)
刘和吴(2018)	(S)
Ruchansky 等人。 (2017)	(S)
舒等。 (2019a)	(大)
我们的工作	(S)

ticles)在分类任务的异构信息网络中相互交互,它们不能应用于归纳设置,即检测新推文的真实性。

基于混合的方法考虑并融合有关源推文的多模态上下文信息。 CSI (Ruchansky et al., 2017)通过结合响应文本和用户配置文件来学习顺序转发特征,并根据用户的社交互动生成可疑的用户分数。王等。 (2018)开发一个事件对抗神经网络来学习可迁移特征

通过删除特定于事件的特征,以及卷积神经网络来提取文本

和视觉特征。 dEFEND (Shu et al., 2019a)联合学习回复评论的顺序效应和新闻内容之间的相关性

和评论,并使用注意力机制来提供可解释性。

我们在表1 中比较了我们的工作和最相关的研究。我们工作的独特之处在于:针对短文本,不需要用户响应评论,并允许模型可解释性。

3 问题陈述

设Ψ = {s1, s2...s|Ψ|}是一组推文故事， U = {u1, u2...u|U|}是一组用户。每个si ∈ Ψ是一个短文本文档（也称为1 , qi 2 , ..., qi
源推文），由si = {q dicating li words 李}在 in story si给出。与用户向量相关联 每个uj ∈ U是 xj ∈ R具有d 维的用户特征。当发布新闻故事d 代表 si时,一些用户将分享si并生成一系列转发记录,这称为传播路径。给定 a,我们将其传播路径表示为新闻故事si , Ri = {...,(uj , xj , tj), ...},其中(uj , xj , tj)表示第 j 个用户uj （和他们的特征向量xj ）

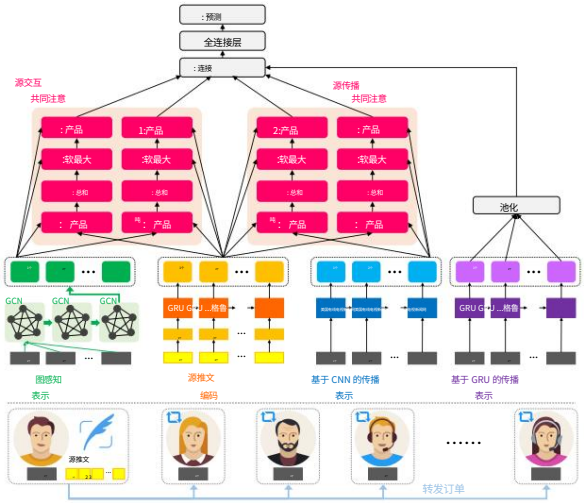

图 1:我们的 GCAN 模型的架构。

谁转发了故事si ,并且j = 1, 2, ..., K （即， K = |Ri |）。我们将转发故事si的用户集表示为Ui 。

在Ri中,我们将在时间t1最初共享si的用户表示为u1 。对于j > 1,用户uj在tj (tj > t1)转发了si 。每个故事si都与一个二进制标签yi ∈ {0, 1}相关联以表示其真实性,其中yi = 0表示故事si是真实的， yi = 1 表示si是假的。

给定源推文si以及包含转发si的用户uj及其特征向量xj的相应传播路径Ri ，我们的目标是预测故事si的真实性yi ,即二元分类。此外,我们要求我们的模型突出显示很少的用户uj ∈ U转发si和很少的单词q ∈ si可以解释k为什么si被识别为真或假。

我

4 提出的 GCAN 模型

我们开发了一种新模型,即图形感知共同注意网络 (GCAN),以根据源推文及其基于传播的用户来预测假新闻。 GCAN 由五个部分组成。首先是用户特征提取:创建特征来量化用户如何参与在线社交网络。第二个是新故事编码:在源推文中生成单词的表示。第三个是用户传播表示:建模和表示用户如何使用提取的特征来传播源推文。第四种是双重共同注意机制:捕获源推文与用户交互/传播之间的相关性。最后是进行预测:通过连接所有学习到的表示来生成检测结果。

4.1 用户特征提取

为了描述用户如何参与社交网络,我们使用他们的元数据和配置文件来定义每个用户uj的特征向量xj 。

提取的特征如下：(1)用户自我描述的字数，(2) uj的网名字数，(3)关注uj的用户数，(4)用户数uj关注的用户数,(5) 为uj创建的故事数，(6) uj的第一个故事后经过的时间,(7) uj帐户是否经过验证,(8) uj是否允许地理-空间定位,(9)源推文的发布时间和uj的转发时间之间的时间差,以及 (10) uj和源推文之间的转发路径长度（如果uj转发源推文则为 1）。最终,生成每个用户特征,其中v是数字

特征向量xj ∈ R。 v

4.2 源推文编码给定的源推文由字级编码

器表示。输入是故事si中每个单词的单热向量。由于每个源故事的长度都不同,我们在这里通过设置最大长度m 来执行零填充。

令E = [e1, e2, ..., em] ∈ R为源故事的输入向量,其中em是第m 个词的单热编码。我们创建一个全连接层来生成词嵌入， V = [v1, v2, ..., vm] ∈ R d×m,其中d是词嵌入的维数。 V 的推导由下式给出：

$$V = \tanh(WwE + bw) \quad (1)$$

其中Ww是可学习权重矩阵， bc是偏置项。然后,我们利用门控循环单元 (GRU) (Chung et al., 2014)从V 中学习单词序列表示。源推文表示学习可以表示为： st = GRU(vt), t ∈ {1, ..., m},其中m是GRU 维度。我们表示源推文

表示为 S = [s , 秒 2 , ..., s m] ∈ R d×m。

4.3 用户传播表示

随着时间的推移,源推文si的传播由一系列用户触发。我们的目标是利用提取的用户特征向量xj以及用户序列传播si来学习用户传播表示。潜在的想法是真实新闻中的用户特征

传播与假传播不同。

我们利用门控循环单元 (GRU)和卷积神经网络 (CNN) 来学习传播表示。

这里的输入是用户转发si的特征向量序列,用PF(si) = xn 表示，其中n是观察到的转发的固定长度。如果共享si的用户数大于n,我 x1, x2, ..., xt , ..., 们取前n个用户。如果数字小于n,我们对PF(si)中的用户重新采样,直到它的长度等于 n。

基于 GRU 的表示。给定特征向量序列PF(si) =我们利用 GRU 来学习传播表示。每个 GRU 状态都有两个输入， ..., xt ,...,, 当前特征向量xt和前一状态的输出向量ht－1,以及一个输出向量ht 。

基于 GRU 的表示学习可以表示为： ht = GRU(xt), t ∈ {1, ..., n},其中n是GRU 的维数。我们生成最终的基于 GRU 的用户传播嵌入h ∈ R erage pooling,由 h = CNN-based Representation给出。我们利用一维卷积神经网络的优势来学习P F(si) 中用户特 d 通过 av征的顺序相关性。我们一次考虑λ 个 新 台币=1 恩。

连续用户来模拟他们的顺序相关性,即xt , ..., xt+λ－1。因此过滤器设置为

Wf∈R _ λ×v . 然后输出表示向量C ∈ R d×(t+λ－1)由下式给出

$$C = \mathrm{ReLU}(Wf \cdot Xt{:}t+\lambda-1 + bf) \quad (2)$$

其中Wf是可学习参数的矩阵，
ReLU是激活函数， Xt:t+λ－1描述了第一行索引来自t = 1 的子矩阵

到 t = n － λ + 1， bf是偏置项。

4.4 图感知传播表示

我们的目标是创建一个图表来模拟转发源的用户之间的潜在交互。这个想法是故事si之间存在某种相关性。具有特定特征的用户可以 = (Ui , Ei) 揭示源推文是假的可能性。为了满足这样的想法， ful为共享源故事si (即Ui)的用户集合构造一个图G ,其中Ei是相应的边集。

由于用户之间真正的交互是非
已知,我们认为 G 是全连接图,即∀eαβ ∈ Ei , uα ∈ Ui , uβ ∈ Ui , uα = uβ, n×(n－1) 2
|伊| = ———— . 为了在图中合并用户特征,每条边eαβ ∈ Ei与

权重ωαβ,权重基于用户特征向量xα和xβ 之间的余弦相似度导出,由ωαβ =给出。我们使用矩阵A = [ωαβ] ∈ R图 G 中的任意一对节点uα和uβ基于源推文si的构造图G创建图卷积网络 (GCN) 层（Kipf和 Welling，2017）。 xα·xβ / xαxβ n×n 代表之间的权重 .

GCN 是一种多层神经网络,它对图数据执行操作并根据节点的邻域生成节点的嵌入向量。 GCN 可以从节点的直接和间接捕获信息

通过逐层卷积堆叠邻居。

给定图G的矩阵A， G中用户的特征向量矩，和X描绘阵 g 维节点特征矩阵H(l+1) ∈ R n×g可以导出为 , 新的

$$H(l+1) = \rho(AH \quad (l)Wl), \tag{3}$$

其中l是层数， A~ = D－归一化对称权重矩阵(Dii = $\frac{}{2}$ AD－ 2是 Aij)， Wl ∈ R d×g是第 l 个 GCN 层的可学习参数矩阵。 ρ是一 j 个激活函数,即 ReLU ρ(x) = max(0, x)。

这里H(0)设置为X。我们选择堆叠两个GCN 层来推导学习到的图形感知表示,表示为 G ∈ R g×n .

4.5 双重共同注意机制

我们认为可以通过调查哪些类型的转发用户关注源故事的哪些部分来揭示假新闻的证据,而虚假线索可以通过转发用户之间的互动方式反映出来。因此,我们开发了一种双重共同注意机制来模拟源推文之间的相互影响（即， S = [s gation embeddings（即,来自第 4.3 节的C = [c ） ）,以及源推文和图形感知之间的相互影响交互嵌入（即G = [g

秒, , ..., s m])和用户propa c n－λ+1]

, , ..., ,

1 2 克, ..., 克 n]来自第4.4 节）。配备共同注意力学习,我们的模型能够通过查看传播中的转推用户与源推文中的单词之间的注意力权重来进行解释。换句话说,通过扩展共同注意公式（Lu et al., 2016）,所提出的双重共同注意机制旨在同时关注源推文单词和图形感知交互用户（源交互共同注意） , 并且还要注意 source-tweet words 和传播的用户同时（source-propagation co-attention）。

源交互共同关注。我们先计算邻近矩阵F ∈ R tanh(S WsgG),其中Wsg是 如： F = 可学习参数的d × g矩阵。通过将邻近矩阵视为一个特征,我们可以学习预测源和交互注意力图,由下式给出

$$\begin{aligned} Hs &= \tanh(WsS + (WgG)F) \\ Hg &= \tanh(WgG + (WsS)F) \end{aligned} \tag{4}$$

其中Ws ∈ Rk×d ,Wg ∈ Rk×g是可学习参数的矩阵。邻近矩阵F可以被认为是将用户交互注意力空间转换为源故事单词注意力空间,反之亦然。然后我们可以通过 softmax 函数生成源单词和交互用户的注意力权重：

$$\begin{aligned} \text{一个} &= \text{softmax}(w\ hsHs) = \\ \text{个} &= \text{softmax}(w\ hgHg) \end{aligned} \tag{5}$$

哪里_ 秒 ∈ R 1×米 g ∈ R _ 1×n 是vec tors of attention probability for each word in the source story and each user in the interaction are 图,分别。 whs, whg ∈ R权值。最终,我们可以 1×k learn 使用派生的注意力权重通过加权和生成源故事词和交互用户的注意力向量,由下式给出

$$\hat{}s1 = \sum_{我=1}^{米} \text{和我} \text{程}, \quad g\hat{} = \sum_{j=1}^{n} \text{个} \text{jjg} \tag{6}$$

其中^s1 ∈ R 1×d和g^ ∈ R 1×g是学习到的共同注意特征向量,描述了源推文中的单词如何被相互交互的用户所关注。

源传播共同关注。分别为源故事和用户传播生成共同注意特征向量^s2 ∈ R的过程与源交互共同注意相同,即创建另一个邻近矩阵以将它们转换到彼此的空间.由于页数限制,我们跳过重复的细节。 1×深 1×深

请注意,基于 GRU 的用户表示不用于学习与源推文的交互。原因是转发序列中的用户个人资料看起来也很重要,正如CRNN （Liu 和 Wu， 2018)所建议的那样,并且应该

表 2:两个 Twitter 数据集的统计数据。

	推特15	推特16
# 源推文	742	412
# 真的	372	205
# 伪造的	370	207
# 用户	190,868	115,036
平均每个故事平均转推。每个来源的单词	292.19 13.25	308.70 12.81

单独强调。然而,基于 CNN的用户表示（即描述用户配置文件序列的特征)已被用于共同注意机制以学习他们的交互

与源推文。

4.6 进行预测

我们的目标是使用源交互共同注意特征向量^s1和g^、源传播特征向量^s2和c^以及顺序传播特征向量h 来预测假新闻。

令f = [^s1, g^, ^s2, c^, h]然后被送入多层前馈神经网络,最终预测标签。我们生成二元预测向量y^ = [^y0, y^1],其中y^0和y^1分别表示标签为0和1的预测概率。它可以通过

$$\hat{y} = \text{softmax}(\text{ReLU}(fW_f + b_f)), \quad (7)$$

其中Wf是可学习参数的矩阵， bf是偏置项。设计损失函数以最小化交叉熵值：

$$L(\Theta) = -y \log(\hat{y}1) - (1 - y) \log(1 - \hat{y}0) \quad (8)$$

其中Θ表示整个神经网络中的所有可学习参数。我们选择 Adam 优化器来学习Θ ,因为它可以中止地确定学习率。

5 实验

我们进行实验来回答三个问题： (1)与最先进的方法相比，我们的 GCAN 模型是否能够实现令人满意的假新闻检测性能？ (2) GCAN 的每个组件对性能有何贡献？ (3) GCAN 能否生成令人信服的解释来强调为什么推文是假的？

5.1 数据集和评估设置数据。 Ma 等人编制的两个著名数据集。 (2017)， Twitter15 和 Twitter16,被利用。每个数据集包含源的集合推文,以及相应的转推用户序列。我们只选择“真”和“假”标签作为基本事实。由于原始数据不包含用户资料,我们使用用户ID通过Twitter API爬取用户信息。

竞争方法。我们将我们的 GCAN与最先进的方法和一些基线进行比较,如下所列。 (1) DTC (Castillo et al., 2011):一种基于决策树的模型,结合了用户配置文件和源推文。 (2) SVM-TS (Ma et al., 2015):一种线性支持向量机分类器,利用源推文和转发用户配置文件的序列。 (3) mGRU (Ma et al., 2016):一种改进的用于谣言检测的门控循环单元模型,它从转发的用户资料中学习时间模式,以及源的特征。 (4) RFC (Kwon et al., 2017):一种扩展的随机森林模型,结合了转推用户配置文件和源推文的特征。 (5) CSI (Ruchansky et al., 2017):一种最先进的假新闻检测模型,结合文章和传播假新闻的用户的群体行为,使用 LSTM 并计算用户分数。 (6) tCNN (Yang et al., 2018):一种改进的卷积神经网络,结合源推文特征学习用户配置文件序列的局部变化。 (7) CRNN （Liu 和 Wu， 2018）:一种最先进的联合 CNN 和 RNN 模型,可以学习转发用户个人资料的局部和全局变化,以及资源推文。 (8) dEFEND (Shu et al., 2019a):一种最先进的基于共同注意力的假新闻检测技术学习源文章的句子和用户配置文件之间的相关性的模型。

模型配置。我们的模型被称为“GCAN”。为了检查我们的图感知表示的有效性,我们创建了另一个版本“GCAN-G”,表示我们的模型没有图卷积部分。对于我们的模型和竞争方法,我们将训练 epoch 的数量设置为 50。GCAN 的超参数设置为:转发用户数 = 40,词嵌入 dim = 32,GRU 输出 dim = 32,1-D CNN输出滤波器大小 = 3,一维 CNN 输出暗淡 = 32,GCN 输出暗淡 = 32。竞争方法的超参数按照各自研究中提到的设置进行设置。

指标和设置。评估指标包括准确性、精确度、召回率和 F1。我们随机选择 70% 的数据用于训练,30% 的数据用于测试。重复进行的训练测试 20

表 3:主要结果。最佳模型和最佳竞争对手分别以粗体和下划线突出显示。

	推特15				推特16			
方法	F1	记录	前	加速器	F1	记录	前	加速器
故障码	0.4948	0.4806	0.4963	0.4949	0.5616	0.5369	0.5753	0.5612
支持向量机	0.5190	0.5186	0.5195	0.5195	0.6915	0.6910	0.6928	0.6932
mGRU	0.5104	0.5148	0.5145	0.5547	0.5563	0.5618	0.5603	0.6612
征求意见稿	0.4642	0.5302	0.5718	0.5385	0.6275	0.6587	0.7315	0.6620
tCNN	0.6200	0.6262	0.6248	0.7374	0.5249	0.5305	0.5296	0.5919
神经网络	0.7174	0.6867	0.6991	0.6987	0.6304	0.6309	0.6321	0.6612
犯罪现场调查	0.6311	0.6384	0.6365	0.7016	GCAN-G	0.7938	0.7990	0.7959
保卫	0.7939	GCAN	0.8250	0.8295	0.8257	0.8767	0.7593	0.7632
18.1%	18.7%	19.3%	15.9%	3.8%	19.9%			

0.5140	0.5206	0.5199	0.5881					
0.6367	0.6433	0.6419	0.7576					
0.6612	0.6541	0.6611	0.6584	0.7383				
0.8636	0.6754	0.6802	0.6785					
0.7594	0.9084	改进	15.0%	20.8%				

次,并报告平均值。

5.2 实验结果

主要结果。主要结果示于Ta

图3. 我们可以清楚地发现,所提出的 GCAN在两个数据集的所有指标上明显优于最佳竞争方法,在 Twitter15 和 Twitter16 中的平均性能分别提高了 17% 和 15% 左右。即使没有提出的图形感知表示,GCAN-G 也可以在 Twitter15 和 Twitter ter16 中将最佳竞争方法平均分别提高 14% 和 3% 。这种有希望的结果证明了 GCAN 对假新闻检测的有效性。结果还暗示了三个见解。首先， GCAN在 Twitter15 和 Twitter16 上分别比 GCAN-G 提高了 3.5% 和 13% 。这展示了图形感知表示的有用性。其次,GCAN 中的双重共同注意机制非常强大,因为它明显优于最先进的非共同注意模型 CSI 。第三,虽然 GCAN-G 和 dE FEND 都是基于共同注意的,但从GCAN-G中的转发用户序列中学习到的额外序列特征可以显着提高性能。

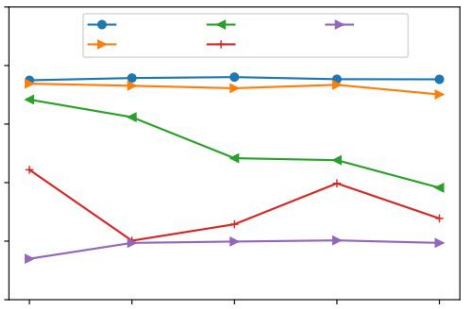

图 2:Twitter15 中 # 转推用户的准确性。

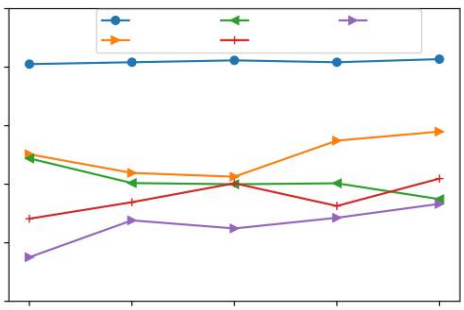

图 3:Twitter16 中 # 转推用户的准确性。

早期发现。我们通过改变每个源故事观察到的转推用户数量（从10到50)进一步报告性能（仅由于页面限制而导致的准确性） ,如图 2 和图 3 所示。显然可以发现我们的GCAN 始终如一地显着优于竞争对手。即使只有十个转发者,GCAN 仍然可以达到 90% 的准确率。这样的结果告诉 GCAN 能够对传播的假新闻进行准确的早期检测,这很重要

在捍卫错误信息时很重要。

消融分析。我们通过从整个模型中移除每个组件来报告每个 GCAN 组件如何做出贡献。下面的“ALL”表示使用 GCAN 的所有组件。通过去除双重共同注意、基于 GRU 的表示、图形感知表示和基于 CNN 的表示,我们有子模型“-A”、“-R”、“-G”、

er15 Twitter16
0.52 0.64
0.59 0.65
.735 0.7
0.88 0.78
0.89 0.88
.915 0.91

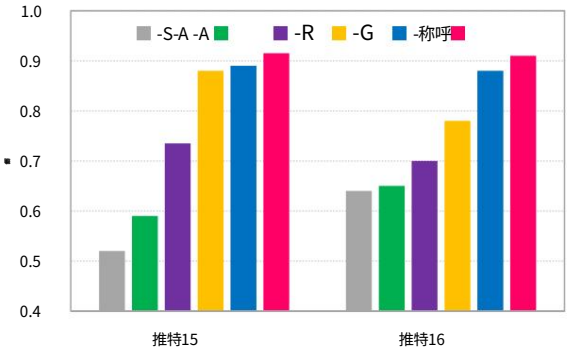

图 4:Accuracy 中的 GCAN 消融分析。

图 5:通过词云突出显示证据词。
较大的字体大小表示较高的共同注意权重。

和“-C”,分别。子模型“-SA”表示没有源推文嵌入和双重共同注意的模型。结果如图4所示。我们可以发现每个组件确实都发挥了重要作用,特别是对于双重共同注意(“-A”)和用户传播和交互的表示学习(“-R”和“-G”)。由于源推文提供了基本线索,因此没有它(“-SA”),准确性会显着下降。

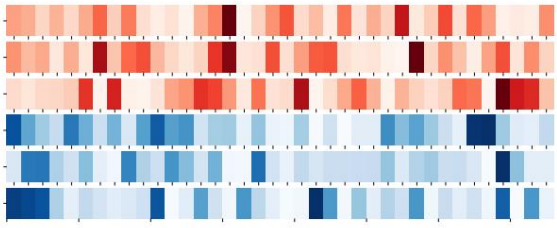

图 6:用户传播 3 个虚假(上 F1-F3)和 3 个真实源推文的注意力权重可视化。从左到右是转发顺序。深色表示较高的注意力权重。

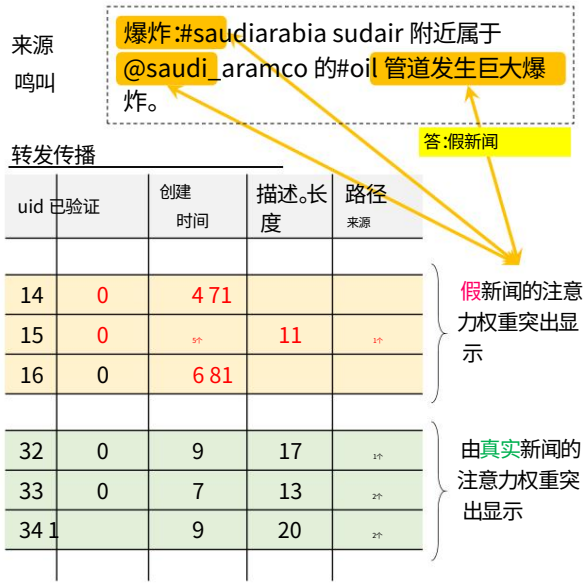

图 7:GCAN 在源推文中突出显示的证据词(上)和 GCAN 在转推传播中突出显示的可疑用户(下),其中每一列都是一个用户特征。请注意,仅提供了很少的用户特征。

5.3 GCAN 可解释性

源自第4.5节的共同关注权重关注源推文单词和转推用户(源传播共同关注)使我们的 GCAN具有可解释性。通过展示注意力权重分布的位置,可以揭示预测假新闻的证据词和用户。

请注意,我们不考虑可解释性的源交互共同关注,因为从构造图中学习的用户交互特征不能直观地解释。

源词的可解释性。为了证明可解释性,我们在测试数据中选择了两条源推文。一个是假的(“breaking: ks patient is risk for epola: in strict isolation at ku med center in kansas city #kwch12”),另一个是真实的(“确认:这无关紧要。rt @ksdknews:确认:#mike-brown 没有犯罪记录。 #弗格森”)。我们在词云的字体大小中突出显示具有较高共同注意权重的证据词,如图 5 所示。GCAN 预测前者是伪造的,对词“breaking”和“strict”的关注度更高,并将后者检测为真实因为它包含“已确认”和“ir

相关的。”这样的结果可能符合常识(Rashkin et al., 2017; Horne and Adali, 2017),即假新闻倾向于使用戏剧性和晦涩的词语,而真实新闻则伴随着证实和事实核查相关的词语。

转推传播的可解释性。我们的目标是利用传播中的转发顺序来揭示假新闻和真新闻之间的行为差异。我们随机选择三个假的(F1-F3)和三个真实的(T1-T3)源故事,并绘制它们

如图 6 所示,从源传播共同注意（第4.5节)中提取权重,其中从左到右的水平方向表示转推的顺序。结果表明,确定

一个故事是否是假的,首先应该考察早期转发源故事的用户的特征。假新闻在用户特征方面的证据可能在传播过程中均匀分布。

转发器特性的可解释性。
我们的 GCAN 模型的源传播共同关注可以进一步提供一种解释,以揭示可疑用户的特征和他们关注的词。图 7 展示了一个案例研究。

我们可以发现,转推传播中可疑用户的特征可以是:帐户未验证、帐户创建时间较短、用户描述长度较短以及到发布源推文的用户的图路径长度较短。此外,他们关注度最高的是“破”和“流水线”两个词。我们认为这种解释有助于解释假新闻的检测,从而了解他们的潜在立场。

六,结论

在这项研究中,我们提出了一种新颖的假新闻检测方法,图形感知共同注意网络（GCAN）。 GCAN 能够根据转发者的顺序预测一条短文本推文是否是假的。问题场景比现有研究更现实和更具挑战性。评估结果显示了GCAN强大的有效性和合理的可解释性。此外,GCAN还可以提供假新闻的早期检测,性能令人满意。我们相信 GCAN不仅可以用于假新闻检测,还可以用于社交媒体上的其他短文本分类任务,例如情绪检测、仇恨言论检测和推文流行度预测。我们将在未来的工作中探索模型泛化。此外,虽然假新闻通常针对某些事件,但我们还将扩展 GCAN 以研究如何删除特定于事件的特征,以进一步提高性能和可解释性。

致谢

这项工作得到台湾科学技术部 (MOST) 的资助,109-2636-E-006-017（MOST 青年学者奖学金)和 108-2218-E-006-036,以及中央研究院在授予 AS-TP-107-M05。

参考

Hunt Allcott 和 Matthew Gentzkow。 2017. 2016 年大选中的社交媒体和假新闻。经济展望杂志,31:211-235。

Carlos Castillo、Marcelo Mendoza 和 Barbara Poblete。 2011. 推特上的信息可信度。在第 20 届万维网国际会议论文集中,WWW 11,第 675–684 页。

Meeyoung Cha、Wei Gao 和 Cheng-Te Li。 2020. 检测社交媒体中的假新闻:亚太视角。公社。美国计算机学会,63(4):68–71。

Junyoung Chung、Caglar Gulcehre、KyungHyun Cho 和 Yoshua Bengio。 2014. 门控递归神经网络对序列建模的实证评估。

Ming-Han Feng、Chin-Chi Hsu、Cheng-Te Li、Mi Yen Yeh 和 Shou-De Lin。 2019. Marine:具有关系邻近性和节点属性的多关系网络嵌入。在万维网会议上,WWW 19,第 470-479 页。

Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. 利用情绪检测社交媒体上的假新闻。 CoRR,abs/1903.01728。

本杰明霍恩和西贝尔阿达利。 2017. This just in:假新闻在标题中包含很多内容,在正文中使用更简单、重复的内容,比真实新闻更像讽刺。在 AAAI 网络和社交媒体国际会议记录中,第 759-766 页。

Jyun-Yu Jiang、Cheng-Te Li、Yian Chen 和 Wei Wang。 2018. 识别在线流媒体服务中共享帐户背后的用户。在第 41 届国际 ACM SIGIR 信息检索研发会议上,SIGIR 18,第 65-74 页。

Thomas N. Kipf 和 Max Welling。 2017. 图卷积网络的半监督分类。在第五届国际学习代表会议论文集,ICLR 17。

Haewoon Kwak、Changhyun Lee、Hosung Park 和 Sue Moon。 2010. 什么是推特、社交网络或新闻媒体?在第 19 届万维网国际会议论文集中,WWW 10,第 591-600 页。

Sejeong Kwon、Meeyoung Cha 和 Kyomin Jung。 2017. 不同时间窗口的谣言检测。 PLOS ONE,12(1):1-19。

Cheng-Te Li、Yu-Jen Lin 和 Mi-Yen Yeh。 2018. 预测社交网络信息传播参与者及其应用。信息科学,422:432 – 446。

杨柳和吴以芳。 2018. 通过循环和卷积网络的传播路径分类早期检测社交媒体上的假新闻。在 AAAI 人工智能会议上,第 254-261 页。

Jiasen Lu、Jianwei Yang、Dhruv Batra 和 Devi Parikh。 2016. 用于视觉问答的分层问题图像共同注意。在第 30 届国际神经信息处理系统会议论文集中,NIPS 16,第 289-297 页。

Jing Ma、Wei Gao、Prasenjit Mitra、Sejeong Kwon、Bernard J. Jansen、Kam Fai Wong 和 Meeyoung Cha。 2016. 用递归神经网络检测微博谣言。 IJCAI 人工智能国际联合会议,第 3818-3824 页。

Jing Ma、Wei Gao、Zhongyu Wei、Yueming Lu 和 Kam-Fai Wong。 2015. 使用微博网站上社会背景信息的时间序列检测谣言。在第 24 届 ACM 国际信息和知识管理会议论文集中,CIKM 15,第 1751–1754 页。

Jing Ma、Wei Gao 和 Kam Fai Wong。 2017. 通过内核学习使用传播结构检测微博帖子中的谣言。在 ACL 2017 - 第 55 届计算语言学协会年会上,会议记录,第 708-717 页。

Jing Ma、Wei Gao 和 Kam-Fai Wong。 2018. 使用树结构递归神经网络在推特上检测谣言。在第 56 届计算语言学协会年会会议记录中,第 1980-1989 页。

卡夏波帕特。 2017. 评估网络声明的可信度。在第 26 届万维网伴侣国际会议记录中,WWW 17 伴侣,第 735-739 页。

Martin Potthast、Johannes Kiesel、Kevin Reinartz、Janek Bevendorff 和 Benno Stein。 2018. 对超党派和假新闻的文体度量调查。在计算语言学协会第 56 届年会会议记录中,ACL 18,第 231-240 页。

汉娜·拉什金、Eunsol Choi、Jin Yea Jang、Svitlana Volkova 和 Yejin Choi。 2017. 深浅不一的真相:分析假新闻和政治事实核查中的语言。在 2017 年自然语言处理经验方法会议记录中,第 2931-2937 页。

Julio CS Reis、Andre Corréia、Fabr´ cio Murai、Adri ano Veloso 和 Fabr´ cio Benevenuto。 2019. 用于假新闻检测的可解释机器学习。
在第 10 届 ACM 网络科学会议论文集中,WebSci 19,第 17-26 页。

Natali Ruchansky、Sungyong Seo 和 Yan Liu。 2017. CSI:一种用于假新闻检测的混合深度模型。
在 2017 年 ACM 信息和知识管理会议记录中,CIKM 17,第 797-806 页。

Justin Sampson、Fred Morstatter、Liang Wu 和 Huan Liu。 2016. 利用社交媒体中的隐式结构进行紧急谣言检测。在第 25 届 ACM 国际信息和知识管理会议论文集中,CIKM 16,第 2377–2382 页。

Kai Shu、Limeng Cui、Suhang Wang、Dongwon Lee 和 Huan Liu。 2019a.捍卫:可解释的假新闻检测。在第 25 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集,KDD 19,第 395-405 页。

Kai Shu、Amy Sliva、Suhang Wang、Jiliang Tang 和 Huan Liu。 2017. 社交媒体上的假新闻检测:数据挖掘视角。
SIGKDD 探索。
Newsl., 19(1):22-36。

Kai Shu、Xinyi Zhou、Suhang Wang、Reza Zafarani 和 Huan Liu。 2019b。用户配置文件在假新闻检测中的作用。 CoRR,abs/1904.13355。

王佩琪和李成特。 2019. 通过学习行为感知异构网络嵌入来发现恐怖分子。在第 28 届 ACM 信息和知识管理国际会议论文集中,CIKM 19,第 2097-2100 页。

王亚庆、马丰龙、金志伟、叶远、光绪勋、基什莱贾、鲁肃和高敬。
2018. Eann:用于多模式假新闻检测的事件对抗神经网络。在第 24 届 ACM SIGKDD 知识发现国际会议论文集中&数据挖掘,KDD 18,第 849–857 页。

Rui Yan、Ian EH Yen、Cheng-Te Li、Shiqi Zhao 和 Xiaohua Hu。 2015. 解决社交网络的致命弱点:基于影响传播的语言模型平滑。在第 24 届万维网国际会议论文集中,WWW 15,第 1318–1328 页。

范洋、杨柳、于小慧、杨敏。 2012.
自动检测新浪微博谣言。在 ACM SIGKDD 挖掘数据语义研讨会论文集中,MDS 12。

阳阳、郑雷、张家伟、崔庆才、李周军和 Philip S. Yu。 2018. Ti-cnn:用于假新闻检测的卷积神经网络。

Zhe Zhao、Paul Resnick 和 Qiaozhu Mei。 2015. 询问思想:从询问帖子中早期发现社交媒体中的谣言。在第 24 届万维网国际会议论文集中,WWW 15,第 1395-1405 页。