我是谁？基于深度学习的文本个性检测

会议论文 - 2018年5月

doi: 10.1109/icc.2018.8422105

5位作者，包括。

孙先国

东南大学（中国）

15部著作 **122次**引用

查看简介

沈小军

密苏里大学-堪萨斯城

73篇出版物 **904次**引用

查看简介

我是谁？基于深度学习的文本个性检测

孙祥国，刘波，曹久新，罗俊洲
东南大学计算机科学与工程学院，南京，211189
电子邮件。{ sunxiangguo, bliu, jx.cao, jluo }@seu.edu.cn

沈小军
计算机和工程学院
密苏里大学堪萨斯城分校，美国堪萨斯城，64110
Email:shenx@umkc.edu

***摘要*-近来，基于网络社交网络文本的人格检测引起了越来越多的关注。然而，大多数相关模型都是基于字母、单词或短语，这不足以获得良好的结果。在本文中，我们提出了我们初步的但有趣的和有用的研究结果，表明文本的结构也可以是研究文本中人格检测的一个重要特征。我们提出了一个名为2CLSTM的模型，它是一个双向的LSTM（长短期记忆网络）与CNN（卷积神经网络）的结合，以利用文本的结构检测用户的个性。此外，还提出了一个概念，即 "潜在句群 "来表达抽象的特征组合。基于紧密联系的句子，我们使用我们的模型来捕捉它。据我们所知，大多数相关工作只在一个数据集上进行了实验，这可能不能很好地解释其模型的多功能性。我们在两种不同的数据集上进行了评估，其中包括长文本和短文本。在这两个数据集上的评估都取得了较好的结果，这表明我们的模型可以有效地学习有效的文本结构特征来完成任务。**

***Index Terms*-computational personality, Big Five, deep learning- ing, online social networks, NLP.**

I. 简介

在心理学领域，一个被普遍接受的、最有影响力的描述和衡量一个人的人格程度的模型是大五模型[1]，它由五个特质组成。开放性、自觉性、外向性、协调性和神经质。

随着网络社交网络的发展，如Twitter、新浪微博[1]等，我们很容易获得用户创造的网络社交文本。很多文章[2], [3], [4]认为一个人的性格特征和他们从网络文本记录中观察到的行为之间存在着很强的联系。因此，基于在线社交网络文本的人格检测对于大量的应用具有重要的意义，比如个人识别系统、心理诊断、人力资源管理等等。

然而，从文本中捕捉真正有用的、与用户个性有密切联系的特征，仍然是一项具有挑战性的探索任务。尽管一些研究人员已经进行了全面的研究，并发现了一些有用的特征，如 LIWC (Linguistic Inquiry and Word Count) [5], Mairesse[6], 字母的水平特征[7]，[8]，响应模式[9]等。不幸的是，他们的大多数应用主要依赖于字母、单词或短语的特征，只能得到有限的结果，这意味着有必要探索更有用的特征。

我们注意到，一些研究人员[10]，[11]已经开始使用集合学习来整合多种特征，而不是仅仅依靠一种特征来提高性能。这是我们未来要遵循的一个正确方向。我们的研究经验表明，探索更多的相关特征是极其必要的，这些特征集合起来可以更完整和准确地反映一个人的个性特征和他的原始文本数据之间的关系。事实上，对用户创造的在线文本的结构的研究，还没有像通过传统方式获得的其他特征那样彻底。显然，为了开发一个有效的多特征模型，我们首先需要学习、分析并对每个单一特征进行良好的建模。

在本文中，我们试图推动对网络文本的结构及其与人格的关系的研究。具体来说，我们对三个问题进行了研究：（1）从网络社交网络的文本中发现的结构是否能有效地反映一个人的个性？(2)
如何对文本结构进行适当的定义？(3)
我们如何从网络文本中捕捉这种结构特征？

最近，深度学习方法被引入到人格预测中，并获得了良好的表现。直观地说，卷积神经网络（CNN）试图重构文章的写作过程，而重现神经网络（RNN）则试图通过模仿类似于人类的阅读过程来理解文本。通过整合这两种神经网络，我们提出了2CLSIM模型，一个双向的LSTMs与CNN相连接，来捕捉文本的结构特征。此外，一个概念潜伏句群（LSG），被引入来描述抽象的特征组合，从密切相关的句子。此外，据我们所知，大多数相关工作只在一个数据集上进行了实验，这可能不能很好地解释其模型的多功能性。我们在两个相对异质的数据集上进行了评估。在这两个实验中，我们的模型都表现出了较好的效果，这说明我们的模型能够正确地捕捉更高级的结构特征，并且能够有效地检测用户的个性特征。总而言之，我们的贡献是

[1] https://weibo.com/login.php

978-1-5386-3180-5/18/$31.00 ©2018 IEE

这项工作的内容如下。

- 我们引入了潜在句群（LSG）的概念，在句子层面上对文本的结构特征进行建模，并使用CNN来学习这些特征。
- 在CNN和RNN的基础上，我们提出了一个组合神经网络模型来实现我们的任务。实验结果表明，我们的模型优于那些只使用单一RNN或CNN的模型。

本文的其余部分组织如下。我们在第二节中介绍相关工作。然后，我们的模型将在第三节中介绍。实验和评估将在第四节中预先发送。之后，我们在第五节中提出结论和未来工作。

II. 相关作品

大体上，本研究使用的数据类型包括文本、头像[9]、喜欢[12]等。对于文本数据，大多数研究将其作为一种特殊的文本分类任务。尽管近年来在文本分类领域已经取得了很大的进展，但从文本中检测个性仍然处于早期阶段。如前所述，获取适用于个性检测的有用特征仍然是一个具有挑战性的问题。以前的大多数工作主要关注的是字符或词级别的特征。[7]的作者将RNN用于字符级和单词级，以建立单词和句子的分层、矢量表征来进行特征推断。Qiu等人[13]分析了Twitter中参与者的词语和他们的性格之间的关系。他们发现人格特征和推文中的特定词汇之间存在着联系。除了一些特定的词语，许多研究者发现，词语的心理意义也可以反映出他们的特征。其中一种用来表达这种关系的方法被称为LIWC（语言学调查和词数）[14]。刘晓倩和朱婷婷[11]使用LIWC来表示每条推文，并通过DFT（离散傅里叶变换）提取主要部分。然后，他们使用堆叠的自动编码器来实现无监督的特征学习。此外，其他一些研究者也发现了使用从文本中提取的结构来进行个性检测的有效性。Honghao Wei等人[9]使用CNN的1,2,3-grams核来捕捉结构。然而，使用固定大小的核子会对长文本造成限制。由于他们的工作是基于叠加概括，而文本数据只是整体的一部分，所以最终的输出结果不能证明简单的CNN是否有效。Navonil Majumder等人[10]使用3维卷积来学习一篇文章的结构。他们对文章进行了从词级到句级的编码。然后他们试图继续使用CNN来构建基于先前工作的文章向量。然而，当把句子向量聚合到文档向量时，由于不收敛，他们无法继续使用卷积操作。为了应对这个问题，他们使用了最大集合法来代替。

受Siwei Lai el[15]的工作启发，我们将两者结合起来。在本研究中，RNN和CNN比GloVe更能捕捉到结构特征。但有以下改进。首先，与他们的工作不同，我们使用LSTM而不是简单的RNN，这样模型在短文和长文中都能得到更好的表现。第二，我们观察到大多数相关的工作只考虑了字符、单词或句子的特征。我们将结构扩展到LSG，LSG是由相关的句子组衍生出来的。

III. 2CLSTM模型

在这一节中，我们介绍了我们的2CLSTM模型的去尾。III-A小节概述了该模型并介绍了信息流的过程。在这个过程之后，模型的一些重要组成部分将在III-B到III-E的小节中介绍。

A. 模式概述

我们模型的架构如图1所示，包括单词嵌入、双向LSTMs、带有1,2,3-grams核的CNN层和分类。整个模型主要由两部分组成：2LSTMs和CNNLSG。基本上，第一部分着重于集合上下文和提取词的语义特征，而第二部分则试图从句子结构中学习特征。五个特征被分别训练。作为CNNLSG的输出，文档向量将被送入softmax层，得到最终的类别。

一个文本的过程需要以下步骤。首先，每个词被嵌入到词向量中，这将在B节中解释，然后上下文信息被我们模型的第一部分编码。在我们模型的第二部分（CNNLSG）中，我们使用CNN来学习第一部分产生的结构特征LSG，这将在D部分解释。第二部分（CNNLSG）产生的最终特征向量将被送入softmax来产生最终的人格特征，如E部分所介绍的。

B. 词语嵌入

对于单词嵌入，我们使用了来自GloVe[16]的预训练的单词向量，这是一种用于获得单词向量表示的无监督学习算法。每个词被嵌入到100个维度。如果一个词没有存在于GloVe单词列表中，一个随机数在[0.25, 0.25]中将被分配到整个坐标上。然而，我们强烈建议从你的数据集中训练你自己的词向量。我们没有这样做，因为我们的硬件和数据集有限。你也可以用word2vect模型代替GloVe，对我们的任务来说没有太大区别。[2]

C. 2LSTMs

原始的LSTM（长短时记忆网络）如图2所示。LSTM不是一个只执行简单激活函数的单元，而是有内部自循环以及

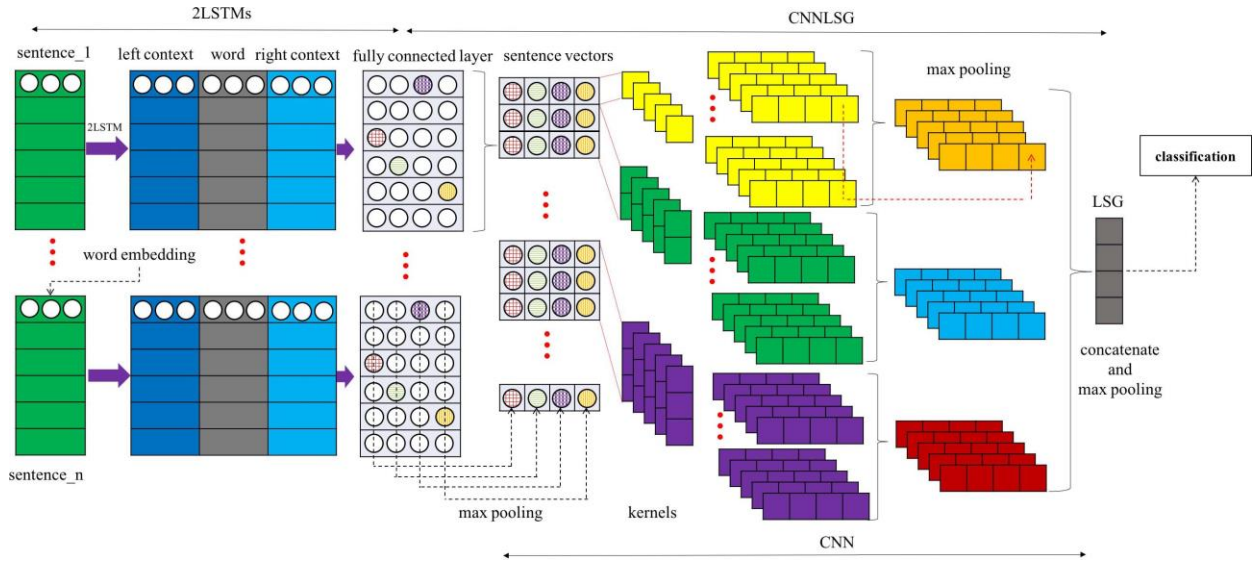


图1: 2CLSTM的结构

RNN的外部递归。每个单元都有更多的参数和闸门来控制信息的流动。

具体来说，LSTM可以在一定程度上容忍单元遗忘左边的上下文知识，这是由遗忘门实现的。信息泄露的百分比通过sigmoid函数设置在0到1之间，如下所示。

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right) \quad (1)$$

其中$\boldsymbol{x}^{(t)}$是输入，而$\boldsymbol{h}^{(t)}$是隐藏层向量。同时，LSTM还通过输入门控制输入知识的吸收，如下所示。

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right) \quad (2)$$

除了使用自己的参数外，它类似于遗忘门。通过在一定程度上过滤掉输入和上下文信息，内部状态被更新如下。

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right) \quad (3)$$

输出$\boldsymbol{h}^{(t)}$，也可以通过输出门关闭，方法如下。

$$h_i^{(t)} = \tanh\left(s_i^{(t)}\right) q_i^{(t)} \quad (4)$$

其中$\boldsymbol{q}^{(t)}$，计算方法与遗忘门和输入门类似。

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \quad (5)$$

为了从上下文中保留更多的信息，我们用双向的LSTMs与当前的单词相连接。其优点之一是，更多来自上下文的结构特征（包括左文和右文）集中在

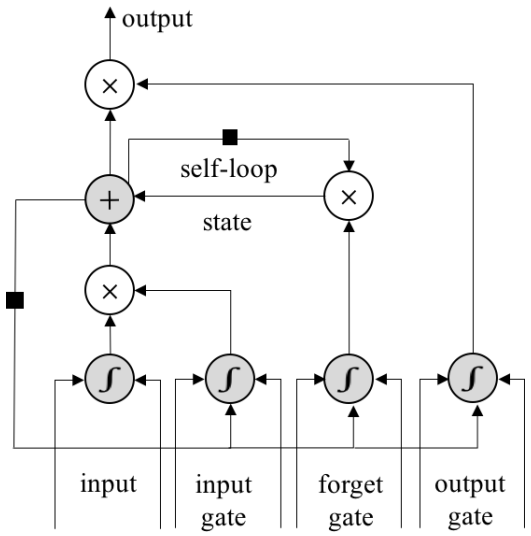


图2：LSTM单元

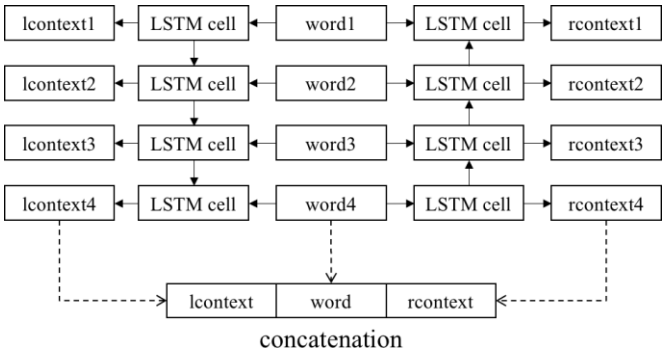


图3：2LSTM的结构

词，这为后续的进一步特征提取奠定了基础。我们将这种结构命名为2LSTMs，如图3所示。

D. 潜伏判决组

句群的概念来自于语言学。基本上，句群是指几个连续的句子，它们在逻辑和语义结构上紧密相连，如坐标关系、偏好关系、因果关系等。然而，对于大多数文本任务来说，检测这些具体的关系是不现实的。事实上，我们通常关注的是句子向量之间在某些方面的关系。这些句子关系是基于一些特定的向量维度，这意味着是一种 "潜在的"关系。因此，我们用潜伏句子组（LSG）这个短语来表示句子特征的抽象组合。LSG意味着从维度的角度来看，几个有密切关系的句子。LSG中的这些句子不需要在空间位置上严格连续。LSG的定义如下。

*定义1（潜伏句群）。*潜伏句群（LSG）被定义为由一些句子向量组成的综合体，这些句子向量在某些坐标上是紧密相连的。

我们使用CNN来学习LSG特征。我们模型的这一部分的细节显示在图1中。通过对每个句子的最大集合，我们得到句子向量。然后在每个维度上，我们使用1,2,3-gram kernels来学习每个坐标上的LSG特征。之后，紧接着是密集层和最大池化层，最终的向量将被产生并送入分类器。

E. 种类

在分类部分可以使用各种分类器，如softmax、SVM、KNN甚至是Adaboost。在大多数深度学习模型中，softmax是使用最广泛的分类器之一。Softmax接收特征向量作为输入，并计算每个类别的概率，如下所示。

$$p(y = i|x;\theta) = \frac{e^{\theta_i^T x}}{\sum_j e^{\theta_j^T x}}$$

其中 y 为类，x为输入向量。获得较高分数的类是softmax的输出。在我们的模型中，我们使用softmax作为最终分类。

IV. 实验

在这一节中，我们提出了一些实验，以评估我们的人格检测模型的有效性。

A. 数据集

我们在实验中使用了以下数据集。

*1) 意识流作文。*我们使用了James Pennebaker和Laura King的意识流文章数据集[17]。它包含了2,467篇有效的匿名文章，并标有作者的人格特征。EXT（外向性），NEU（神经质），AGR（合意性），CON（自觉性）和OPN（开放性）。

*2）YouTube。*YouTube的个性数据集[3]，来自大约400个YouTube博客的网络视频。它包含了从视频中手动转录的语音、性别和行为特征。与第一种数据集的一个主要区别是，在这个数据集中大多数文本都比较短。另一个区别是，这里的标签（个性印象）是由注释者通过观看每个视频收集的印象，而不是由作者自己收集的。

备注。我们选择这两个数据集是因为我们希望测试我们的模型是否能够处理不同的情况。表一显示了我们从这两类数据集中总结出来的一些统计数据。我们选择这两个数据集的主要原因如下。

- 首先，如图4所示，大多数来自YouTube的文档都比意识流文章短，这可以验证2CLSTM模型是否对短文和长文都有效。
- 其次，意识流文章数据的标签来自于作者自己的问卷调查，这可以解释为自我认知，而另一个数据集（YouTube数据集）可以被视为外在认知，因为这个数据集的个性标签来自于观看微博作者视频的志愿者。因此，这两个数据集可以证明我们的模型在两种情况下都是有效的，无论标签是由作者还是由其他人产生。

表I:数据集的宏观统计

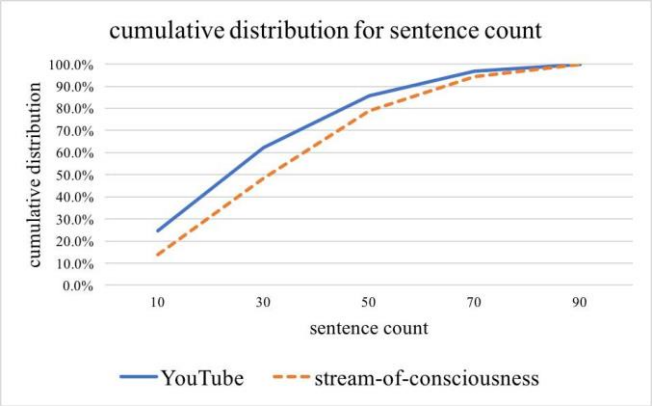
财产	意识流	youtube
平均字数	648	526
最大字数	2488	1972
最小字数	33	36
平均句数	46	41
最大句数	327	147
最小句数	1	2

B. 对比模型

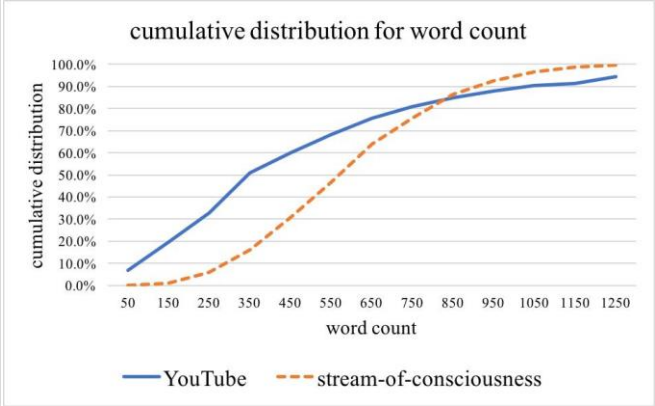
在本小节中，我们选择了五个著名的模型作为我们的对比模型，它们被列在表三中。

第一种是基于TF-IDF feature的贝叶斯分类法，这是最基本和最常见的文本分类方法之一。TF-IDF是术语频率（TF）和反文档频率（IDF）的产物。该值显示了一些特定词对区分文档的重要性。然而，在实践中，TF-IDF值太小（例如几乎为零）或太高的词是没有必要考虑的。在我们的实验中，我们设定TF-IDF的有效范围为[0.2, 0.5]。

[3] https://www.idiap.ch/dataset/youtube-personality



(a)



句 (b)字数

图4：YouTube和意识流数据集的句子和单词的累积分布情况

我们使用的第二个对比模型是2CNN，即来自Honghao Wei等人[9]的二维卷积网络。由于文本数据只是他们整个数据集的一部分，而且该任务只是他们异质集合中的一个分类器，我们从他们的CNN中删除了LIWC特征和保留的文档向量。我们选择它作为我们的对比模型之一，因为这个模型是最新的方法，它专注于单词之间的关系。

第三个模型是3维CNN，由Navonil Majumder等人提出[10]。他们使用3维的卷积核（词向量、短语和句子）。由于计算量很大，我们将句子合并为一个，然后使用二维卷积。所有来自1,2,3-gram kernels的映射都用零填充，以保持相同的形状。最后，我们再次串联这些映射，并使用最大池化层来获得文档向量。我们选择它的原因是，这个模型是最新的方法，它试图基于句子来学习特征。

至于最后两个对比模型，它们都属于RNN方法。一个是单一的RNN（1LSTM），另一个是没有当前单词连接的双向LSTMs。它们都是文本分类中的典型模型。

C. 实验环境

我们的实验基于ubuntu 16.04 LTS，8GB内存和Intel Core i7-4790 CPU。一个文档中的最大字数设定为800（如果小于它就垫零），字数向量为100维。根据这些，每个层的输出形状如表二所示。此外，我们为研究目的开放了我们的源代码[4]。

作为普通操作，我们将数据集分成测试数据和训练数据（大约9 : 1）。在每个训练历时中，大约20%的训练数据被当作验证数据。这里设定的最大历时次数是100，但在实际实验中，大多数任务可以提前在20-30个历时时停止。为了克服过拟合问题，我们增加了一些剔除层

表二：主要层的结构形状

层数	输出形状
双向LSTM	(800, 600)
与当前词串联	(800, 700)
带有1-gram的Conv2D	(25, 300, 10)
带有2-gram的Conv2D	(24, 300, 10)
带有3-gram的Conv2D	(23, 300, 10)
储存量最大的2D	(3, 300, 10)
储存量最大的2D	(1, 300, 1)
润彩客网	(2)

*

为了克服过拟合，我们增加了一些下降率为0.2至0.3的下降层，这些层没有在表中列出。

掉落率从0.2到0.3，这取决于情况。

D. 结果

表中列出了这些模型对意识流数据集和YouTube数据集的宏观精度。

三。我们用粗体字来标记每个类别中的前两个最佳结果。总的来说，我们可以发现2CLSTM模型在所有类别中的表现都相当好。虽然在某些类别中，如意识流数据集的CON或OPN，以及YouTube数据集的AGR或OPN，2CLSTM并没有名列前茅，但它仍然保持了相对较好的精度。此外，对于意识流数据集，2CLSTM在AGR和NEU等类别中得到了更好的结果，而对于YouTube数据集，它在EXT和NEU中得到了更好的结果。

YouTube的宏观精度通常优于意识流，这在一定程度上反映了这里列出的模型（包括2CLSTM）似乎更擅长于探测他人眼中的人格特征比作者本人。

[4] 源代码可用于研究目的。你可以从这个网址下载：https://github.com/sunxiangguo/2CLSTM