利用深度学习甄别高质量留言

张馨瑞
（西北师范大学附属中学，甘肃兰州，730070）

摘要：在Web1.0时代，用户与互联网交互方式仅限于单向只读，进入Web2.0后，由于计算机运算和存储能力显著提高，用户主导网站生成内容变得越发便利，如何在良莠不齐的海量留言中挑选高质量文本，是一个亟待解决并充满实际意义的文本分类任务。本文利用近些年发展迅速的神经网络技术并且结合电商平台亚马逊公开的推荐数据集，设计一套不需要先验知识就可甄别高质量留言的神经网络模型。

关键词：文本分类；神经网络；亚马逊推荐数据集

DOI:10.16589/j.cnki.cn11-3571/tn.2020.20.035

1 概述

■ 1.1 研究背景与意义

近几年，互联网技术得到飞速发展，相关应用产品如雨后春笋般层出不穷，例如淘宝、今日头条、亚马逊商城等。用户在使用这些应用的同时，会自发进行留言和评论，这些文本不仅可以丰富、补充被评论目标，而且可以更加客观的帮助用户了解被评论目标的质量和特点，使用户感受到更加人性化的服务。互联网用户基数庞大，但是由于教育背景、所处环境、生活地域等因素限制，用户对目标事物的见解和表达具有局限性，局限性大的文本对其他用户帮助很小，我们将其定义为低质量文本。反之，定义为高质量文本。如果能够快速挑选出高质量用户留言（评论）并将其优先、公开地展示，用户通过浏览质量高的留言信息，就能够在第一时间更了解目标事物，这样的使用经历无疑会提升用户满足感，直接或间接影响应用收益。

■ 1.2 研究现状

当前甄别高质量文本的方法可分为两种：基于“众包”的投票机制和基于人工特征的机器学习模型。两种机制都具有比较明显的缺陷：“众包”机制面临的主要问题是留言价值的人工标注数据分布不均衡，具体表现在畅销及热门的目标会吸引大量的用户参与留言以及对留言的价值做出评价，但是对于一款新上架或者冷门的消费品，短时间内很难通过“众包”的方式快速收集有关留言价值的人工标注（投票），通常畅销或者热门的消费品仅仅只占全部的一小部分，绝大多数消费品分布在不被关注的“长尾”；基于人工特征的机器学习模型在很大程度上依赖大量的语言学分析工具：如中文分词、词性标注、句法分析器等，增加了模型训练结果的不确定性，首先，这些语言学工具无法保证对目标语言分析的准确性，可能在模型训练阶段引入偏差；其次，针对多种语言的应用，对应需要对多种语言学工具进行开发或购买，将会花费大量的人力和金钱成本；最后，究竟哪些语言学特征有助于不同场景下的高质量文本的筛选，相关研究人员众说纷纭，没有最终定论，如果针对每个场景进行特定特征开发，既不现实也无法保证最终效果。

■ 1.3 研究内容

本文使用神经网络中常用的词向量、卷积神经网和循环神经网络等技术完成一套可以甄别高质量商品留言的神经网络模型，需要的数据集可由电商平台亚马逊推荐数据集中的商品描述、用户留言和用户投票构建所得。用 AUC 值评估训练完成后的模型，通过评测结果判断模型甄别高质量留言的能力。

2 相关技术

词向量（Word Embedding）是指通过构建语言模型对自然语言进行表征学习的技术，可以将词表中单词映射为对应的数字向量。具体地，它是将每个单词所占一维的空间映射到维度更低的连续向量空间。本文使用 Word2Vec 官方文档中经过 Skip-Gram 方法预训练的词向量完成后续任务，使用该方法得到的词向量有两个明显优势：第一，向量维度较低，其长度通常在 50 至 300 之间，有效摆脱了维度灾难困境，便于深层神经网络的进一步计算；第二，有较好的向量空间意义，词义和用法相似的词语通常聚集在一起，不同词向量在空间中的距离可较为客观的反应出不同词语间的差异性。基于上述优点，词向量在深度学习中扮演着中举足轻重的角色，是保证很多自然语言任务高效完成的默认前提。

卷积神经网络（Convolution Neural Network, CNN）可通过卷积提取局部信息特征，利用参数共享策略减少网络中的参数规模，起到纵向加深网络层数作用。一般情况下，深层网络与浅层网络相比，前者能够更好拟合数据分布，也能更高效地学到特征和目标任务间的数学关系。本文使用 CNN 提取英文词汇中的字符级别特征。

循环神经网络（Recurrent Neural Network, RNN）用于提取时序对象中的关联特征，在自然语言处理和语音任务中被广泛使用。普通 RNN 不能有选择的进行信息累积，只能暴力积累过去时刻的所有信息，这样导致跨度较大的关联特征被短期关联特征不断覆盖。本文使用 RNN 的变体——双向长短期记忆单元（Bi-LSTM）捕获文本序列的语义特征，

LSTM 在普通 RNN 的基础上增加了遗忘门、更新门和输出门的运算结构，有效的解决了循环神经网络随着文本序列不断增长，无法捕捉相隔太远信息的设计缺陷。

2014 年 Bandana 等人在机器翻译任务中使用了注意力机制并取得优异结果。自此之后，注意力机制成为深度学习中发展最快的方法之一，在特别在诸如机器翻译、阅读理解和文本生成等自然语言处理任务中大放异彩。人类在浏览文本信息时，在每个词汇上投入的精力并不相等，而是有重点的进行浏览理解，注意力机制便是基于这样的人类行为设计而来，让模型能够从海量信息挑选出数据规模小，但是内容重要的关键信息。本文使用注意力机制对商品描述和用户留言中的语义特征进行提取以及融合。

3 数据处理和模型实现

■ 3.1 数据处理

亚马逊的推荐数据集内包含多个文件，商品描述和用户留言分别在两个不同的文件中，由于它们包含相同的“商品 ID”属性，本文通过关联操作将两者合二为一，处理后的数据集包含商品 ID、商品描述、用户留言和用户投票四个关键字，其中每个商品 ID 对应唯一的商品描述和多个对应的用户留言和用户投票，数据的抽象格式和实例如下所示。

抽象数据格式：

{

ID: Product ID Number,

Product: Product Description,

Review: [{Text: Review_1, Helpful: [Help Vote, Total Vote]},

{Text: Review_1, Helpful: [Help Vote, Total Vote]}, …]

}

实例（亚马逊平台销售的 surface Pro 6 商品）：

{

ID: EX806001,

Product: ‘Unplug, pack light. Get productive your way, all day. Wherever you are, new surface Pro 6 makes it easy to work and play virtually anywhere…’,

Review: [{Text: ‘Same day shipping and received on time. The laptop is so elegant. Light weight, and the battery life last more than a day…’, Helpful: [77, 104]}, …]

}

在上述数据集中，每条用户留言中都存在唯一的用户投票（Helpful），它代表“众包”投票的统计结果。具体地，其中第一字段 Help Vote 表示认为该条留言对自己有帮助的用户人数，第二字段表示浏览过该条留言的总人数。根据这两字段的具体含义，对高质量留言进行人工定义，具体公式如下：

$$p = help\_vote / total\_vote \tag{1}$$

$$y = \begin{cases} 1 & if \quad p \geqslant 0.6 \\ 0 & if \quad others \end{cases} \tag{2}$$

公式 1 中的 p 某条留言对其他用户有帮助的概率值，公式 2 表示如果概率值大于 0.6，即超过 3/5 的用户对该条留言内容持肯定态度，此时将标签值 y 定义为 1，表示高质量留言；反之，标记为 0 的 y 值表示低质量留言。

■ 3.2 模型实现

用户在电商平台购物时，为进一步了解目标商品，经常会浏览与商品相关的留言信息，

在海量留言中发掘对自己有帮助的信息，该行为通常会花费大量的用户时间，降低用户在平台的消费体验。针对该情况，本文基于亚马逊公开的推荐数据集，设计了一套端到端可甄别高质量留言的神经网络模型。模型分为将自然语言映射为向量表示的混合词向量模块、能够捕获商品描述和用户留言语义特征的编码模块和对齐并融合商品描述与用户留言信息的注意力模块。模型整体基于 TensorFlow（TF）深度学习框架实现，具体结构如图 1 所示。

3.2.1 混合词向量模块

混合词向量（Hybrid Word Embedding）模块是将词向量和基于相应词汇获取的字粒度特征向量拼接形成的混合向量表示。本文选择 Word2Vec 官方文档经过 Skip-Gram 方法预训练并且维度为 50 的词向量对自然语言进行

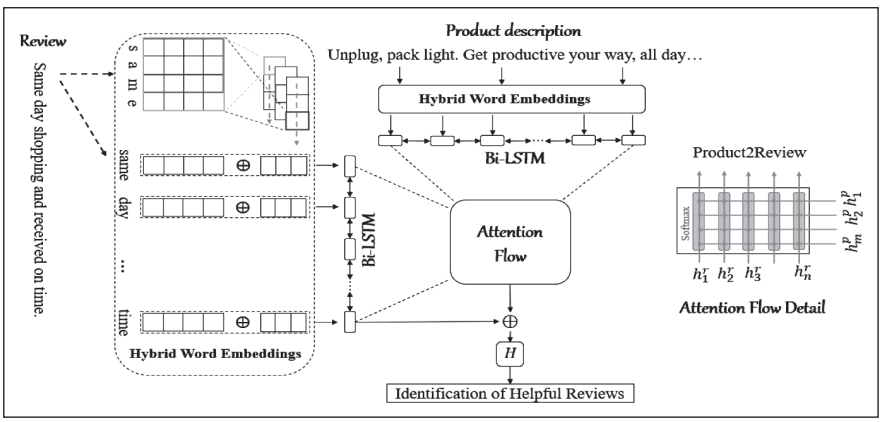

图 1　甄别高质量留言网络模型结构示意图

初始化，字符级别特征则通过字向量和 CNN 共同计算所得。接下来，详细介绍获取混合词向量的具体过程：首先，对商品描述和用户留言进行分词，$R=[r_1,...,r_n]$ 和 $P=[p_1,...,p_m]$ 分别表示分词后的留言与商品描述序列，这里遵循自然语言任务中的常规做法，去掉所有序列中的停用词与低频词，有效缩小词表规模，进而提高词向量的查找速率；然后，利用 CNN 提取单词序列中的字符级别特征，卷积核 $W\in R^{h\times k}$ 学习的随机初始化参数，h 表示卷积核宽度（依次使用 1,2,3 作为卷积宽度），k 表示卷积核长度（长度和字向量维度相等，50），最大池化每个卷积操作获取的特征图，获得维度为 3 的字粒度特征向量；最后，将词向量和单词对应的字符级别特征向量进行横向拼接，即可得到 53 维的混合词向量。利用混合词向量初始化 R 与 P 序列，获得对应的混合向量序列 $\vec{R}=[\vec{r}_1,...,\vec{r}_n]$ 和 $\vec{P}=[\vec{p}_1,...,\vec{p}_n]$。

3.2.2 捕获语义特征模块

在该模块中，利用双向的长短记忆期单元（Bi-LSTM）捕获商品描述和用户留言在句子层级的语义特征，计算过程如公式（3）所示：

$$\vec{H}^R=Bi-LSTM\left(\vec{R}\right),\vec{H}^P=Bi-LSTM\left(\vec{P}\right) \tag{3}$$

其中 $\vec{H}^R$ 和 $\vec{H}^P$ 分别表示用户留言和商品描述经过 Bi-LSTM 隐藏层的输出值，两者的序列长度保持不变，但向量维度增至 106 维，用符号 2d 表示。

3.2.3 流式注意力模块

2018 年，Minjoon 等人针对阅读理解任务提出了 BIDAF 模型，原文使用流式注意力对问题文本和阅读材料进行信息融合。本文参考上述机制中的部分结构，完成商品描述和用户留言间的信息融合。首先，利用向量序列 $\vec{R}$ 和 $\vec{P}$ 计算两者间的注意力概率分布矩阵，具体计算过程如下所示：

$$\vec{S}_{ij}=\alpha\left(\vec{H}^R_{:t},\vec{H}^P_{:j}\right)\in R \tag{4}$$

$$\alpha\left(\vec{h}^R;\vec{h}^P\right)=W^T\left[h\ u\right] \tag{5}$$

公式 5 中的 $\vec{h}^R$ 和 $\vec{h}^P$ 分别与公式 4 中的 $\vec{H}^R_{:t}$ 和 $\vec{H}^P_{:j}$ 等价，分别表示 $\vec{R}$ 中的第 t 列向量和 $\vec{P}$ 中的第 j 列向量，$W\in R^{4d}$ 为随机初始化的可学习参数，符号 d 表示词向量和字粒度特征向量的维度之和，即 d 代表数字 53；然后，利用注意力矩阵 S 计算融合了相关商品描述信息的用户留言向量表示，运算过程如下所示：

$$a_t=softmax\left(\vec{S}_{t:}\right)\in R^m \tag{6}$$

$$\vec{H}^{\bar{R}}_{:t}=\sum_j a_{tj}U_{:j}\in R^{2d} \tag{7}$$

公式 6 利用 SoftMax 函数对 S 进行归一化，m 表示商品描述的文本长度。公式 7 表示对注意力概率值和相应商品描述向量序列进行加权平均运算，其结果即为携带商品描述信息的用户留言向量表示。公式 8 表示对用户留言向量和融合商品描述的留言向量进行多样性的横向拼接：

$$\vec{H}=\left[\vec{H}^R;\vec{H}^{\bar{R}};\vec{H}^R\odot\vec{H}^{\bar{R}}\right]\in R^{6d\times n} \tag{8}$$

具体地，符号 ; 表示拼接操作，运算符 ⊙ 逐点相乘计算，输出向量 $\vec{H}$ 已经包含了商品描述和相关留言具有的多级语义特征，使用该向量对高质量留言进行最终预测，预测结果用 $\tilde{y}$ 表示，具体公式如下所示：

$$\tilde{y}=\text{Sigmoid}\left(\sum_{t\in n}W_p^T\vec{H}_t\right) \tag{9}$$

公式 9 利用 Sigmoid 函数对留言质量进行判别，其中 $W_p\in R^{6d}$ 表示可训练的随机初始化参数。自此，模型结构全部介绍完成。本文利用 TensorFlow 框架完成模型的自动求导和反向传播，具体地，模型的损失函数为二分类的交叉熵，选择随机梯度下降算法作为模型的优化方法，经过多轮训练，不断减小预测值 $\tilde{y}$ 和真实值 y 间的差异，当损失值收敛并且小于人工阈值时，模型训练完毕。

4 实验结果

本文使用 AUC 评估模型效果，而没有使用一般分类任务常用的准确率进行评测，具体原因如下：高质量留言是基于“众包”投票结果获得的，以 0.6 的阈值对文本质量进行划分的行为过于粗狂，可能引入行为噪音，所以在模型的测试阶段，如果仅使用单一且明确的阈值衡量模型稳定性，无法对模型做出真实评估。

本文分别使用亚马逊数据集中电子和工具两个数据量大的类别对模型进行训练和预测，其中训练集和测试集是根据 8/2 原则划分得来的。电子类和工具类的 AUC 评估值分别是 0.611 和 0.604，比基于传统机器学习的基线模型普遍高出 0.2 个点左右。这样的实验结果表明，本文提出的模型能够较好的甄别高质量留言，且模型本身设计简单，易于实现。

参考文献

* [1] Miao Fan, Yue Feng, Mingming Sun, et al. Multi-task neural learning architecture for end-to-end identification of helpful reviews[C]. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 343–350.

* [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and tra

（下转第 91 页）

作，这样的工作需要涉及到电路、电器等，因此需要员工的专业水平较高并且对上岗条件有着严格的要求。此外，在智能化技术落实前，全部都是由相关人员运用现有的公式在手工设计方案的基础上进行估计、推导，这样可能会存有较大的误差。而现在将智能化技术运用到电气工程后，运用计算机中的软件与智能化技术相结合而形成的方案设计，将会大幅度使实验设计时产生的人为误差减少，从而将设计的精准度提升，同时科学合理的运用计算机程序可减少设计过程所使用的时间，从而使工作效率提高[2]。

现代我国为了追求经济飞速发展的同时还格外注重经济发展的效益与质量，把科技转换为生产力便是科技进步最为关键的一步，在生产过程中添加最新科技技术研究成果，从而将整个行业的实力提升。针对电力行业来讲，这将直接影响到我国经济未来发展以及人民生活、活动以及生产是否可以顺利进行。因此在促进电力行业未来发展中，应该将先进技术的作用发挥出来。将现代计算机技术与电气工程相结合，通过运用机器来代替人工，在节省人力资源的同时还可以最大程度使工作效率提高，从而减少大量认为事务，实现经济效益最大化。

■ 2.3 关于 PLC 应用以及神经网络应用

关于神经网络的子系统共有两部分，第一便是电气动态的参数可以对定子电流进行控制以及辨别，而第二便是机电系统的参数可以针对转子速度进行控制以及辨别。针对这样的系统来说，较为常见的算法之一便是反向学习，其不但可以对负载转矩的变化和非初始速度进行管控，还可以将定位时间大幅度缩减。在这样的系统中智能函数估计器具有较强的抗噪音性以及一致性，不再使用控制模型，所以在信号处理以及模式识别等方面，智能神经网络的运用范围较大，可以有效控制电气传动。通常运用尝试等方法可以对出现的问题进行解决。运用反向传播技术可以得到非线性函数相近值，并在最短的时间内得到结果，这将对网络节点有着直接的影响，在进行网络权重调整时，只需要将误差反馈即可[3]。

关于 PLC 的应用，目前我国科技正在飞速发展中，针对电力生产方面也在不断提升其工作要求。针对电力系统，在当前环境下，继电控制器将被 PLC 系统所代替。这种系统主要用途有：可以运用在工艺控制中、可以对企业生产的流程进行协调。针对煤能源运输主要构成有：上煤、储煤、配煤以及辅助系统。主要经过传感器、远程 I/O 以及主站层进行操作，从而提升生产效率，让电气工程中加入 PLC 系统会有很多优点，例如可以进行自动切换等。本文所写如图 1 所示。

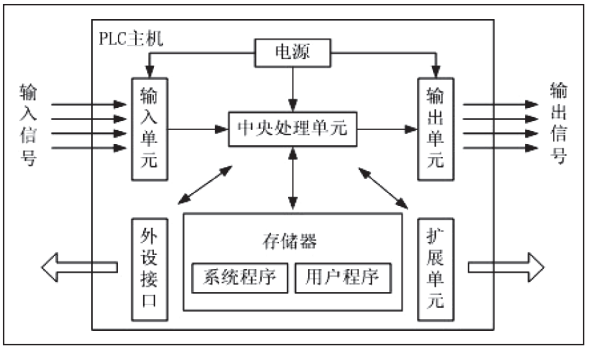

图 1　PLC 工作原理

3 结论

综上所述，电气工程项目在积极推动我国经济发展方面起到非常重要的作用，虽然电气工程项目控制方面在最近几年已经飞速发展，但是仍然存有部分问题。在电气工程项目系统中添加智能科学技术，除了可以使电气设备效率提高，还可以将因为设备而造成的工程损失减少，为推动我国未来电气工程健康发展打下坚实的基础。

参考文献

＊ [1] 姜丽梅 . 智能化技术在电气工程自动化控制中的应用分析 [J]. 电子世界 ,2020(14):56-57.

＊ [2] 肖萍 . 智能化技术在电气工程自动化中的应用 [J]. 科技资讯 ,2020,18(17):31+33.

＊ [3] 宋文强 . 智能化技术在电气工程自动化控制中的应用 [J]. 通信电源技术 ,2020,37(08):67-69.

（上接第 85 页）

nslate[J]. arXiv preprint arXiv:1409.0473, 2014.

＊ [3] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, et al. Bidirectional Attention Flow for Machine Comprehension[J]. In Proc. ICLR 2017.

＊ [4] Yoon Kim. Convolutional Neural Networks for Sentence Classification[C]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 1746–1751.

＊ [5] R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering[J]. In Proc. WWW, 2016.

＊ [6] Tomas Mikolov, Kai Chen, Greg Corrado, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv:1301.3781, 2013.