专栏：信息安全

基于 Bert 模型的互联网不良信息检测

蔡鑫

（中国电信股份有限公司研究院，上海 200122）

摘　要：针对互联网不良信息检测这一业务场景，探讨了基于网站文本内容进行检测的方法。回顾了经典的文本分析技术，重点介绍了 Bert 模型的关键技术特点及其两种不同用法。详细描述了利用其中的特征提取方法，进行网站不良信息检测的具体实施方案，并且与传统的 TF-IDF 模型以及 word2vec+LSTM 模型进行了对比验证，证实了这一方法的有效性。

关键词：不良信息；Bert 模型；文本分析；特征提取

中图分类号：TP393

文献标识码：A

doi: 10.11959/j.issn.1000−0801.2020303

Internet bad information detection based on Bert model

CAI Xin

Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: In view of the business scenario of bad information detection on the internet, the method of detection based on the text content of the website was discussed . Classical text analysis techniques were reviewed. The key technical features and two different usages of Bert model were introduced. The specific implementation scheme of using the feature extraction method to detect website bad information was described in detail, and was compared with the traditional TF-IDF model and word2vec+LSTM model. The validity of this method is verified.

Key words: bad information, Bert model, text analysis, feature extraction

1　引言

互联网是人们获取信息的一个重要媒介。互联网能够不受空间限制进行信息交换，扩展了人们的交流方式，开阔了人们的视野，丰富了人们的知识。但是，在互联网上也存在一些不良的信息内容，比较普遍的就是一些黄赌毒内容。这些不良信息，一方面就像精神鸦片，会毒害和侵蚀青少年的成长，也会让很多普通人沉溺于低级趣味；另一方面，有这类内容的网站，往往会架设在国外的一些云主机或服务器上。当国内用户访问的时候，就会产生大量的关口局跨境流量，不仅占据了出口带宽资源，也造成了运营商大量的结算费用支出。

在传统的方式里，可以通过用户众包模式，例如有奖举报，再配合大量的人工审核，比如色情网站的鉴黄师，去维护一个所谓的黑名单库。然后由运营商对黑名单库中的 URL 进行拦截，达

收稿日期：2020−09−10；修回日期：2020−11−01

到阻断不良信息内容的目的。但是这类网站往往也会通过不断变换域名、更新网页背景图，或者更新部分文字内容等方式翻新，躲避黑名单库的审查过滤，而且这种行为实施起来也是快速但廉价的。

因此，运营商希望能在网络流量中，借助智能算法自动识别那些新出现的包含不良信息内容的URL。一般而言，网站中包含的内容有文本、图片、视/音频等类型，不同的智能算法可以基于这 3 类信息源进行识别，本文主要探讨基于网站文本内容的智能检测方法。

2 文本分析技术进展

2.1 TF-IDF 词袋模型

文本智能分析属于自然语言处理（natural language processing，NLP）范畴。在深度学习应用于 NLP 之前，以 TF-IDF（term frequency-inverse document frequency，词频−逆文本频率指数）为代表的词袋模型（bag of words model）是长文本分析中的关键技术。

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术。字词的重要性随着它在该文本中出现的次数呈正比增加，但同时会随着它在全体语料库中出现的频率呈反比下降[1]。

通过 TF-IDF 模型，文档可以表示成以词典作为维度的一个向量形式，在这种向量化表示的基础上，可以通过机器学习的方法构建模型，解决相应的业务问题。可以看到，基于词袋模型的文档表示方法，虽然考虑了词的重要程度，但是只是根据词的统计特性表示一个文档。而没有考虑到词在文中的次序等一些上下文信息。比方说有这样两句话，“姜文/的/哥哥/是/姜武。”“姜文/是/姜武/的/哥哥。”这两句话的 TF-IDF 数据表示形式是一样的。但是它们的语义是截然相反的。

TF-IDF 的特征表示天然带有高维和稀疏的特性。后来也有通过潜在语义分析（latent semantic analysis，LSA）[2]、潜在狄利克雷分布（latent Dirichlet allocation，LDA）[3]等方法，提取文档低维稠密的高阶特征，并在高阶特征的基础上进行机器学习。对于机器学习算法而言，这种降维操作通常是有益的，但是这些方法也都没有改变 TF-IDF 模型上下文语义丢失的问题。

2.2 深度学习模型与词/字向量

在基于深度学习的 NLP 模型中，文档一般表示成词（甚至是字）的一个序列。这样词/字在文档上下文的次序关系信息就能够保留下来。不同于 TF-IDF 模型对于词所采用的高维稀疏的独热编码（one hot representation）方式，深度学习模型对于词或字往往采用相对低维的向量空间编码（distributed representation）方式，一般称为词向量或字向量。

词向量和字向量一般是在海量语料库的基础上，通过无监督学习的方法而获得。其中，语义相近的词或字，它们的词向量和字向量的空间距离就会比较近，这是分布式编码相对于独热编码的重要特征。有一个著名等式可以类比词向量和字向量的算法处理，即：

$$king-man+woman = queen$$

最初的 word2vec 词向量基于 CBOW（continues bag of words）或 SkipGram 模型训练获得[4]，这里有一个比较大的问题，就是这种简单词向量没法解决词的多义问题。也就是说，当一个词在不同的语境环境里有不同语义的时候，简单 word2vec 词向量没法表达这种差异。

近两年词向量预训练模型技术出现了突破性的发展。其中，2018 年 Matthew 等[5]提出的 Elmo 是一个双层的双向 LSTM 模型。Elmo 实现了动态生成语境向量的能力，能够揭示词和句子更深层的语义。在 6 项 NLP 任务中，创造了当时最好的纪录。同年，OpenAI 团队发布 GPT（generative pre-training）模型，率先将基于 self attention 机制的 transform 技术应用到 Encoder-decoder 网络中，让文本的表征充

分学习到不同的语义侧重点[6]。GPT 模型一举刷新了 12 项 NLP 任务中的 9 项榜单，名噪一时。

4 个月后，2018 年 10 月 11 日，谷歌 AI 团队的 Jacob[7]等基于 GPT 模型的思想，改进提出了 Bert 模型，重新改写了 NLP 任务的全部纪录，在 NLP 领域引起前所未有的巨大轰动，成为 NLP 全面走向预训练模式的跨时代的标志。

3　Bert 模型关键技术

3.1　Bert 模型网络结构特点

Bert 模型的网络结构如图 1 所示，它具有以下 4 个主要特征。

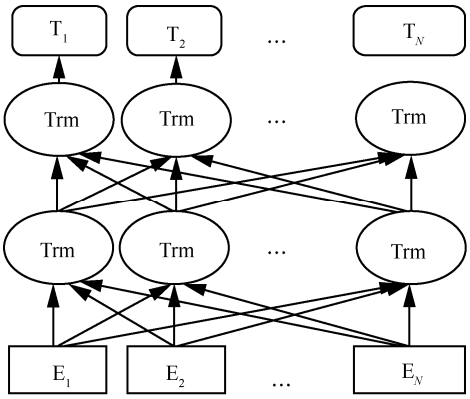

图 1　Bert 模型网络结构

（1）利用真正的双向 Transformer 结构提取上下文关系信息。

（2）利用完形填空式的 Mask-LM（mask-language model）预测被遮挡的词块（token）。

（3）利用下一句预测（next sentence prediction）任务学习句子级别信息。

（4）进一步完善和扩展了通用任务框架，支持更多的场景任务。

另外，Bert 模型创新了词块（token）和句子的表征方式，如图 2 所示。

对于给定词块，Bert 模型的输入表征（input representation）通过对相应的词块嵌入（token embedding）、段嵌入（segment embedding）和位置嵌入（position embedding）的加和构造，从而实现了多语境的差异化表示。而对于单个文本句子或一对文本句子（例如，[问题，答案]），Bert 模型能够通过词块的序列进行明确表征。

3.2　Bert 模型的用法

Bert 的论文中对 Bert 预训练模型设计了两种应用于具体领域任务的用法，一种是微调（fine tune）方法，一种是特征抽取（feature extract）方法。

fine tune 方法指的是加载预训练好的 Bert 模型，以预训练的网络参数权重为蓝本，把下游具体领域任务的数据喂给该模型，在网络上继续反向传播训练，不断调整原有模型的权重，通过端到端的学习训练，从而获得一个适用于新的特定任务的模型，这是一种常见的迁移学习手段。

feature extract（特征提取）方法指的是不再进行编码即嵌入（embedding）的重新训练，直接调用已预训练好的 Bert 模型，对新任务的句子做编

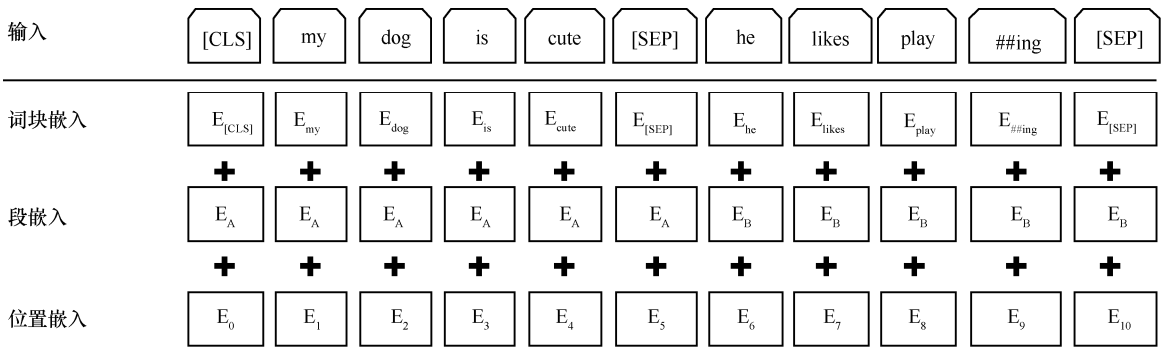

图 2　Bert 模型的输入表征

码，将任意长度的句子编码成定长的向量。具体做法是：将输入序列的一层或多层的隐层状态拼接起来，作为序列的表示和下游模型算法的输入。这种方法没有反向传播过程。

fine tune 的使用方法有一定的限制，需要下游任务对网络的设计要求，与预训练任务具有一致性。feature extract 的方法，不需要再对文本序列的特征表示进行二次训练，形成的句子编码具有更广泛的场景通用性，操作过程比较简便，对于机器性能的要求也会被相对比较低。

因此，本文选择使用 Bert 模型 feature extract 方法，实现对于不良信息网站的检测分析。

4 Bert 模型方法对比实证

4.1 实验设计

本文面对的业务场景，是对给定的网站 URL 进行检测，判别其是否属于涉及赌博题材的网页。根据数据挖掘的一般知识，可以将该问题抽象为一个二分类问题，即根据输入特征利用模型算法将网页分成“是”或“否”涉赌两类。

为了验证 Bert 模型 feature extract 方法的有效性，选取了两种较为常规的文本分类方法作为对照组，分别是：TF-IDF 模型、word2vec+LSTM 模型。

在对比的指标上，选择了在 2 分类问题里常用的 3 个评估指标。

- 精确率（precision）：$P = TP / (TP + FP)$ 。它表示：预测为正的样本中有多少是真正的正样本，是针对预测结果而言的。Precision 又称为查准率。
- 召回率（recall）：$R = TP / (TP + FN)$。它表示：样本中的正例有多少被预测正确了，是针对原始样本而言的。recall 又称为查全率。
- $F1$ 值：$F1 = 2PR/(P+R)$。它表示：精确值和召回率的调和均值，是对上述两个指标的综合评价。

4.2 实验数据准备

根据业务方提供的赌博类网站列表，得到 1 392 个中文内容页面为主的黑名单网站样本。从 Alex topX 网站中选取中文页面网站，再从中进行抽样，得到 1 824 个白名单网站样本。

利用爬虫工具将黑白样本网站的首页 HTML 内容爬取下来。使用 Python 环境下的 beautiful soup 网页分析工具包，将 HTML 标签、JavaScript 脚本等与实际网页题材内容无关的信息过滤掉，仅保留中文字符和标点作为文本净荷。

将该文本净荷作为待验证的 Bert 模型 feature extract 方法，以及对照组 TF-IDF 模型、word2vec+LSTM 模型的初始数据输入。

4.3 Bert 模型方法实施

（1）环境准备

本文所述 Bert 模型 feature extract 方法依赖于 Bert 中文预训练模型，这是本文方法的重中之重。

此外，Bert as service 作为 GitHub 上的一个高评分第三方项目，对于 Bert 的 feature extract 用法，实现了很好的封装，以 Web 接口的形式提供句子编码服务。本文在实证研究中也引入了该项目作为辅助工具。Bert as service 运行环境要求为：Python>=3.5，Tensorflow>=1.10。

接下来需要部署 Bert as service 的服务器端和客户端工具。安装完成后，启动 Bert as service 服务。

（3）Bert 模型特征抽取

客户端调用 Bert as service 方法，将一个网页文本净荷作为一个句子单位。服务端接收到文本句子之后对句子进行定长编码，并返回客户端，通过这种方式实现黑白样本的文本序列特征抽取。

Bert as service 句子编码定长默认 768 维。

（3）分类模型构建

样本中随机选择一定比例（本文设定比例 80%）作为模型训练集，剩余部分作为独立测试集。分类模型输入数据见表 1。

表 1　分类模型输入数据

对比项	样本比例	样本个数		Bert 模型抽取特征数
		黑样本	白样本	
模型训练集	80%	1 114	1 459	768
独立测试集	20%	278	365	768

在 Python 环境下，选择 XGboost 集成分类算法，以上述 768 维的句子编码作为输入特征，针对训练集数据进行分类模型训练。

4.4　对照组模型方法实施

（1）TF-IDF 模型方法

采用 Python 环境下的 jieba 分词工具，对网页文本净荷进行分词。

利用 gensim 工具包中封装的 TF-IDF 算法，提取网页文本净荷 TF-IDF 统计特征，词典做了适当截断以避免特征的高维问题。将该统计特征作为分类器输入，使用 xgboost 集成分类算法进行分类模型构建。

（2）word2vec+LSTM 模型方法

同样，首先要对网页文本净荷进行分词，可以使用上述同样的工具和方法。

接下来是 word2vec 词向量 embedding 学习以及 LSTM 神经网络搭建。为简化实施过程，使用了模块化的神经网络库 Keras 框架。Keras 充分利用 Tensorflow 通用计算能力，并对词向量 embedding 以及包括 LSTM 在内的各种神经网络单元进行了很好的封装，从而减小编程开销，更专注于深度学习模型本身。

完成分词的网页文本净荷，经过 embedding 向量化，进入 LSTM 层进行上下文学习，之后 LSTM 的输出结果经过全连接的 Dense 层将维度降至目标变量的类别个数（此处为 2），利用 sigmoid 作为激活函数，就可以得到输入文本净荷在两个类别的概率分布，从而完成分类模型构建。

4.5　模型效果评估对比

模型训练好之后，针对 Bert 模型 feature extract 方法、以及对照组中的两个文本分类方法，分别基于同样的独立测试集进行预测和评估。

基于独立测试集的模型效果评估指标数据见表 2。

表 2　试验模型效果评估对照

对比项	本文方法	对照方法	
	Bert feature extract	TF-IDF	word2vec+LSTM
精确率	0.97	0.75	0.95
召回率	0.93	0.71	0.91
*F*1 值	0.95	0.73	0.93

5　实验分析

从对照实验的结果来看，采用 Bert 模型的特征提取方法，与传统的 TF-IDF 模型相比，模型效果各项指标都有了很大的提升，这可能跟 Bert 模型很好地保留了文本的上下文信息有着密切关系。同时说明，特征提取方法对于提取网页文本内容的特征信息是有效的。

与同样保留上下文语义的采用简单词向量 word2vec+LSTM 模型相比，Bert 模型 feature extract 方法的效果评估指标略好，但不明显。初步判断主要由于测试集与训练集语料在领域主题上比较接近，基于训练集所得到的词向量用在测试集上语境恰好合适。训练集和测试集数据取自同一个时间范围，测试集上应该也没有太多未编码的超纲新词。因此，word2vec+ LSTM 模型的效果同样也不错。

从模型训练效率看，Bert 模型特征提取方法由于不需要执行 embedding，因此训练效率比较高。从模型预测执行效率看，Bert 模型特征提取方法需要搭建 Bert as service 服务器环境，还需要通过 Web services 调用获得网页文本净荷的编码，增加了交互步骤和复杂度，此为该方法的一个短处所在。

在一部分从日志系统中提取的待测 URL 中，

使用本文所述方法进行了实际应用测试，对模型的分类结果进行了人工校对，发现与在独立测试集上的评估效果无显著性差异，说明该模型具有稳健性。

6 结束语

基于深度学习的 NLP 技术研究在近几年的快速发展，促使 Bert/GPT/XLnet 等预训练模型在相关 NLP 任务中取得了卓越的表现，也成为实际业务场景中解决文本分析类问题的主要选项。

本文选择了基于 Bert 模型的 feature extract 方法，针对互联网不良信息文本内容检测场景开展了应用验证。通过对模型原理的分析，认为该方法具有通用性，在很多文本分析相关的业务场景都可以作为选项进行试用，再通过与传统分析方法的效果进行比较和权衡，决定取舍。

对于 Bert 模型的另外一种主流工作模式，也就是 fine tune 方法，本文尚未展开具体验证。与此同时，注意到开源社区中一个来自 Huggingface 的 Transformers 项目，提供了包括 BERT、GPT、GPT-2、Transformer-XL、XLNET 在内的 NLP 领域预训练模型的封装和调用框架，为预训练模型使用提供了更好的便利性和操作性，迅速得到了社区高度关注和支持。后续将进一步开展对 Transformers 项目的深度研究和验证，力求探索出 NLP 预训练模型在现网应用的最佳实践模式。

参考文献：

[1] 蔡鑫, 娄京生．基于 LSTM 深度学习模型的中国电信官方微博用户情绪分析[J]. 电信科学, 2017, 33(12): 136-141.
CAI X, LOU J S. Sentiment analysis of telecom official micro-blog users based on LSTM deep learning model[J]. Telecommunications Science, 2017, 33(12): 136-141.

[2] SCOTT D. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990(41): 6.

[3] BLEI D M, NG A Y, JORDAN M I, et al. Latent dirichlet Allocation[J]. Journal of Machine Learning Research, 2012(3): 993-1022.

[4] MIKOLOV T，CHEN K，CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781, 2013.

[5] PETERS M, NEUMAN M, IYYER M, et al. Deep Contextualized Word Representations[J]. arXiv:1802.05365, 2018.

[6] RADFORD A, SALINMANS T. Improving language understanding by generative pre-training[Z]. 2018.

[7] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[作者简介]

蔡鑫（1975− ），男，中国电信股份有限公司上海研究院高级工程师，主要研究方向为数据分析挖掘、人工智能、数据规划和信息安全。