基于 Bert 预训练模型的虚假新闻文本检测

王国泰，董晶晶，高 杨*，王 乾
（北京卫星导航中心，北京 100094）

摘 要：虚假新闻剥夺了人们获取真相的权利，也给社会和国家带来了许多危害。文中以虚假新闻文本为例，分析和验证了多种预训练语言模型在虚假新闻文本分类上的分类效果。经实验证明，相较于其他语言模型，Bert 预训练语言模型取得了较好的结果，预测准确率为 86.97%，召回率为 88.12%，F1 值为 87.54%。

关键词：自然语言处理；Bert；文本分类
中图分类号：TP391.4 文献标识码：A 文章编号：1009－2552(2022)01－0137－06
DOI：10.13274/j.cnki.hdzj.2022.01.025

Fake news text detection based on Bert pre-training model

WANG Guo-tai，DONG Jing-jing，GAO Yang，WANG Qian
（Beijing Satellite Navgation Center，Beijing 100094，China）

Abstract：Fake news deprives people of the right to obtain the truth，and it brings trouble to the society and the country. This paper takes fake news text as an example to analyze and verify the classification effect of various pre-training language models on fake news text classification. Experiments result proves that compared with other language models，the Bert pre-training language model has achieved the best results，showing the 86.97% in prediction accuracy，recall rate of 88.12%，and F1 value of 87.54%.

Key words：natural language processing；Bert；text categorization

0 引 言

近年来，随着网络的迅速发展，互联网中的虚假信息泛滥，严重影响了网络环境，其中，以虚假新闻文本为代表的虚假信息最为常见。由于虚假新闻旨在吸引读者的注意，所以传播速度往往更快，且不易甄别。虚假新闻带来的危害尤其值得重视，它不仅给个人、企业带来了不可避免的麻烦，在国家安全层面上，也给政府造成了一定程度上的冲击。经过调查和研究发现，互联网中的虚假新闻甚至会影响某些国家的选举结果[1]。

作者简介：王国泰(1994－)，男，学士，助理工程师，研究方向为信息技术与数据分析。
* 通讯作者：高杨(1986－)，男，硕士，工程师，研究方向为信息技术与数据分析。

为应对这种现象，美国在 2010 年开始讨论虚假检测技术。在同年的 8 月份，人民网舆情中心开始着手进行虚假信息检测技术的研发。虚假新闻自动化检测方法变得尤为重要。

1 相关工作

虚假新闻文本检测任务的核心在于如何获取文本的向量表示，并提取向量特征。利用预训练的语言模型可以更快速、更准确地帮助我们提取到文本的向量表示。

1.1 Word2vec

Word2vec 是 2013 年由谷歌的 Mikolov 提出的一种基于神经网络的概率语言模型[2]。它能将自然语言中的词语或句子转化为方便计算的稠密向量。在 Word2vec 出现之前，人们通常以 One-hot Encoder(独热编码)的方式来转换自然语

言。但是独热编码的结果通常会导致维度爆炸，向量太过稀疏，常常给后面的计算带来麻烦。而 Word2vec 则很好地解决了这一问题，它使用 Vector Representations 来将独热编码转换为低维度的连续稠密向量，并且能将文本中意思相近的词语映射到向量空间中的同一位置。Word2vec 采用两种相反的模型来规范数据的输入和输出，也就是 CBOW（continuous bag-of-words）模型和 Skip-Gram 模型（如图 1 所示）。在 CBOW 模型中，它以输入的文本为预测输入，目标词为预测结果，在这种模型中学习单词的词向量。而 Skip-Gram 则正好相反，它是以目标词为输入，以预测文本为输出来学习单词的词向量。同时，在这两种模型中，Word2vec 并不在意词之间的连接顺序，在训练结束以后获得每一个词语的词向量，他们可以用来表示词与词之间的关系。

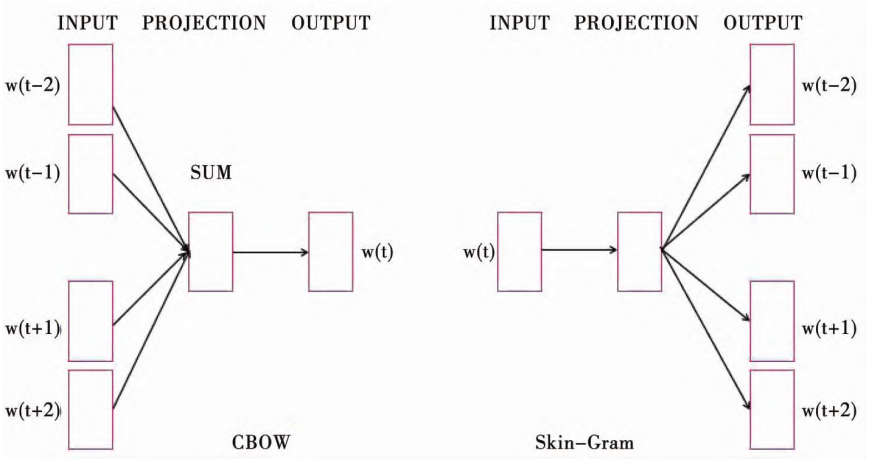

图 1　CBOW 和 Skip-Gram 模型图

1.2　Elmo

Elmo 是一种深层的词向量表示语言模型，它能够充分利用上下文来对单词的复杂特征进行建模，并且能够同时体现多义性（相同词语在不同的上下文中可能具有不同的含义）。Elmo 预训练语言模型在大型文本语料库上进行预训练，在训练过程中通过一个双向的 LSTM 网络来学习单词的词向量表示[3]。Elmo 模型可以用来处理 NLP 中的各种具有挑战性的问题，比如问题问答、文本含义和情感分析等。Elmo 的模型结构如图 2 所示。

Elmo 利用两个前后的 LSTM 网络来实现单词的特征提取。

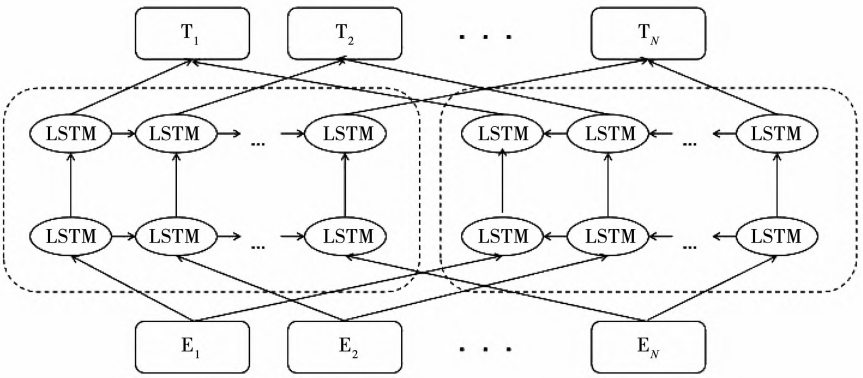

图 2　Elmo 模型结构图

前向 LSTM 结构：

$$p(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}p(t_k\mid t_1,t_2,\cdots,t_{k-1}) \quad (1)$$

后向 LSTM 结构:

$$p(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}p(t_k\mid t_{k+1},t_{k+2},\cdots,t_N) \quad (2)$$

最大似然函数:

$$\sum_{k=1}^{N}(\log p(t_k\mid t_1,t_2,\cdots,t_{k-1})+\log p(t_k\mid t_{k+1},t_{k+2},\cdots,t_N))$$

1.3　OpenAI GPT

OpenAI 在文献 [4] 中提出了 GPT 模型,在文献 [5] 中提出了 GPT2 模型。GPT2 是一个基于 Transformer 结构的语言模型,它在原有的 GPT 上面做了扩展,GPT2 的训练数据包含 40GB 的互联网文本,具有超过 10 倍的参数,数量接近 1.5 亿。GPT2 通过训练可以直接预测本文中的下一个单词。但是由于担心技术被恶意应用,OpenAI 仅仅对外公开发布了一个小的 GPT2 模型。

GPT2 采用两阶段训练法来进行训练。包括第一阶段的无标签文本训练和第二阶段的 fine-turning。在训练时,GPT2 会采用多路注意力机制,这种机制与 Bert 机制相似,但是不同的是 OpenAI GPT2 采用 Mask 机制来对单词的下文进行覆盖遮挡。例如一个句子包括 4 个单词 [A,B,C,D],Mask 机制会读入 A,然后预测后面的 B,在这之后可以看作 AB 都已录入,然后用 AB 来预测 C,并且以此类推。需要注意的是,当我们读入 A 时,BCD 都是被掩盖起来的,在算法里会用一个无穷小来替换被掩盖的值,如图 3 所示。

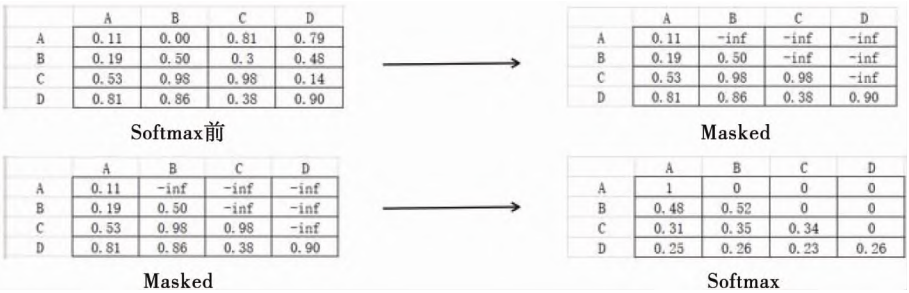

图 3　GPT 的 Mask 流程图

2　Bert 预训练语言模型

Bert 是一个建立在神经网络上的语言处理模型[6]。Bert 模型更加注重于识别句子中单词与单词之间的关系或者是句子与句子之间的关系,它采用一个半监督学习和语言来表示模型,Bert 是一个基于双向 transformer 的模型,它可以共同调节 left-to-right 的 transformer 和 right-to-left 的 transformer。在预训练阶段,Bert 使用无监督的预测任务执行预训练,该任务包括下文遮蔽的语言模型 MLM(Masked Language Model),在执行完预训练后,Bert 模型会针对下游任务进行 fine-turning 来微调模型参数,以达到最适应的效果[7]。

2.1　双向 transformer

自然语言是人类社会中逐渐形成的一种交流方式,一个句子或者一个单词通常需要结合上下文才能体现出它的涵义。这就决定了计算机去理解自然语言不能只是单单的从上文(顺序解析)来解释词,或者是从下文(逆序解析)来解释词,而是要求上下文结合的方式。Bert 的深度双向 transformer 则正是体现了这一理念。

2.2　Masked Language Model

使用双向解释的时候可能会有循环的存在,这就会给单词在“自己”的理解上带来误解。Bert 采用 MLM 模型来解决这种误解[8]。MLM 模型会对输入句子的单词进行随机的 Mask(OpenAI GPT 也是采取相似的这种方式),例如有下面的句子:

“The man went to the store with his dog。”

80% 的概率用“[MASK]”标记来替换——“The man went to [MASK] store with [MASK] dog”。

10% 的概率用随机采样的一个单词来替换——“The man went to A store with B dog”。

10% 的概率不做替换——“The man went to the store with his dog”。

2.3　Embeddings

Bert 模型的词编码不是简单的词编码，而是包含了三层涵义的编码的结合。第一层编码是单词自己的编码，Bert 初始化时会存在一个外部输入的词表来进行编码，这个词表会包含该种自然语言的所有单词。第二层编码则是基于单词的位置信息进行嵌入，为了体现单词在句子中的位置信息，Bert 会将每一个句子中的每一个单词进行 position embedding。第三层编码则是句子级别的编码，为了体现句子的独立性（Bert 称之为 segment embedding），Bert 采用两句话拼接的方式进行构建编码。在三层编码都已经完成后，Bert 再进行三种 embedding 的结合，最后才决定词向量。

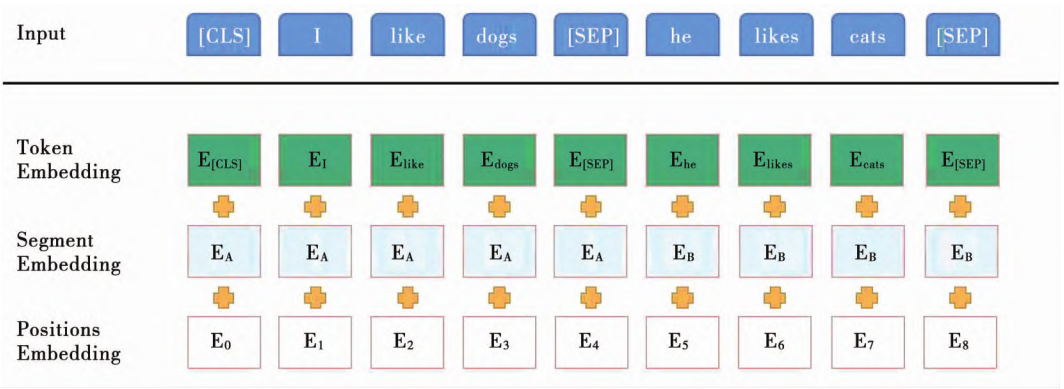

图 4　Bert Embedding

2.4　fine-turning

一个语言模型在训练完成时参数就已经确定了，但是对于不同的下游任务，利用同一套参数显然是不科学的，如果重新训练又会花费大量的时间。Bert 是一个预训练语言模型，为了适应各种不同的下游 NLP 任务，采用了 Fine-turning 的方式来对已经训练好的模型进行微调，以达到最好的模型匹配效果。

3　实验

3.1　预训练数据集

通用的 Bert 模型只采用通用语料库（Wikipedia 和 BooksCorpus）作为预训练数据集。基于虚假文本的特点，为了更好地训练 Bert 的 Fine-tuning 模型，本文从 FakeNewsNet 网站上收集了各种类型的新闻文本数据，如金融新闻、教育新闻、军事新闻等，该数据集包含新闻内容和正确标注真假新闻标签的社会语境特征。表 1 中列举了文中使用到的 Fine-turning 数据集。

①Wikipedia 和 BooksCorpus：原始的预训练数据用于训练原版的 Bert（Wikipedia 总计 13GB 文本，词量大概在 3 亿左右；BooksCorpus 总计 15GB 文本，词量大概在 4 亿左右），它们两者都是通用领域的语料库。

②FakeNewsNet 新闻文本：从 FakeNews 网站上下载的新闻领域的专用语料库（总共包含 26GB 的新闻文本内容，大约 6 亿词量），它涵盖金融、教育、军事等多个领域的虚假新闻。

表 1　语料库数据集

语料库	大小	词量
Wikipedia	13GB	3 亿
BooksCorpus	15GB	4 亿
FakeNewsNet	26GB	6 亿

3.2　实验环境

操作系统为 Ubuntu16.04，CUDA 版本为 8.0，CUDNN 版本为 7.0.5，pytorch 开发工具包为 1.3.1，深度学习框架为 SK-learn 0.21.3。

3.3　实验内容

与 Word2vec 模型、Elmo 模型和 OpenAI GPT 模型在文本分类上作对比，验证 Bert 模型在虚假新闻文本分类分析上的表现。在数据预处理阶段，首先去除文本停用词和出现频率特别高的词，同时去除含有大量的乱码文字或者表情的文本；然后判断是否有缺省值，若有则选择人工填充；接

下来将清洗好的数据集进行划分，本文采用 5 折交叉验证方法，首先将数据集进行 shuffer，得到洗后的随机数据集，然后按 8∶1∶1 划分训练集、测试集和验证集，最后取平均值。在下游任务中，选择 SVM 算法来进行预测，并对比预测结果[9-11]。

3.3.1　预训练

本文使用的预训练模型与 Bert 官方提供的预训练模型相同（包括 BertBase 和 BertLarge），他们具有相同的预训练参数。在训练期间，没有采用 Tensorflow，而是利用 Apache Hadoop YARN 的分布式训练框架，使用 Horovod 库（Uber 基于 Baidu-allreduce 开发的一个开源库）实现了同步和数据并行的分布式训练策略，用一种能够同时调用多个 GPU 的计算架构来训练参数，大大减少了训练时间。Horovod 体系结构使用分布式优化策略，在计算模型权重之前，使用 allreduce 操作来降低平均梯度值。一般情况下，参与计算的 GPU 数量越多，模型的 loss 值就会越低。与 TensorFlow 的架构相比，Horovod 体系架构能够快 3 ~ 4 倍，同时也具有更好的扩展性。

文献 [12] 提出了一种混合精密训练方法，采用这种方法来训练深度神经网络能够大大减少内存切换和计算操作的耗时。本文也尝试利用这种混合精密训练的方法来训练模型。在 FP16（半精度格式）中存储激活函数、梯度和权重比，在 FP32（全精度格式）中进行更新操作，使用损失缩放来保持梯度值的缩小，使用 FP16 算法积累成单精度输出，然后存储到内存。

3.3.2　Fine-turning

Bert 的训练需要经过两个步骤，第一个是预训练，另一个就是 Fine-turning。首先在预训练阶段对大型语料库进行无监督的预训练，然后在 Fine-turning 阶段对下游 NLP 任务进行有监督的微调。在无监督语料库上，从零开始对 F-BERT 进行预训练，并将其微调到下游监督虚假新闻文本检测任务。

①虚假新闻文本边界检测。

虚假新闻文本边界检测是一项基本任务。它能通过消除文本的歧义来检测句子的边界，从文本中分割出正确的句子。

②虚假新闻文本情感分析。

情感分析也是虚假新闻文本检测的任务之一。本文使用一个范围为 [-1,1] 的评分区间来预测每个目标的情绪评分，情绪越趋近于 1，真实率越高，表示该文本为正向情感的新闻文本，反之越低。

3.4　结果分析

评价模型的主要性能指标有：

准确率：

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100\% \tag{3}$$

其中，TP 表示真正例；FP 表示伪正例，准确率表示预测为正例样本中真正正例的样本比例。

召回率：

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% \tag{4}$$

其中，TP 表示真正例；FN 表示伪反例，召回率表示预测为正例的样本占所有正样本的比例。

F1 评测值：

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

不同模型经过训练得到的准确率、召回率以及 F1 值如表 2 所示。经过对相同分类算法下不同的预训练语言模型的比较，OpenAI GPT 和 Bert 预训练语言模型的预测结果明显高于 Word2vec 和 Elmo 模型，由于新闻文本具有语言的时序性，所以具有双向 LSTM 的模型会更有优势。分析可知，Bert 语言模型更有优势，在各方面都取得了最佳性能，准确率达到了 87%，召回率达到了 88%，F1 值高达 87%。

表 2　实验结果

方法	Precision	recall	F1
Word2vec	80.23	85.19	0.8264
Elmo	81.28	82.24	0.8176
OpenAI GPT	84.42	83.74	0.8408
Bert	86.97	88.12	0.8754

4　结束语

本文分析和比较了处理自然语言的一些预训练语言模型，包括 Word2vec 语言模型、Elmo 语言

模型、OpenAI GPT 模型和 Bert 模型，并且针对这四种模型在 SVM 分类算法下对虚假新闻文本的分类预测做了对比实验[13]，实验证明，Bert 预训练语言模型取得了最佳的预测效果。

当然，本文对于虚假新闻文本的研究还有不足之处。例如，在下游的文本分类算法的选择上面，只使用 SVM 算法进行二分类。今后将从多种下游分类算法上面进行纵向的分析比较，找到最佳 Bert 模型的最佳分类算法。

参 考 文 献：

[1] 刘乐. 虚假新闻的危害、成因及治理办法 [J]. 新闻传播，2019(21)：21－22.

[2] 隋浩. 基于 Word2Vec 的微博情感新词识别与倾向判断研究 [D]. 南宁：广西大学，2016.

[3] 张俊遥. 基于深度学习的中文命名实体识别研究 [D]. 北京：北京邮电大学，2019.

[4] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013.

[5] Mikolov T, Sutskever I, Kai C, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26.

[6] 严佩敏，唐婉琪. 基于改进 BERT 的中文文本分类 [J]. 工业控制计算机，2020，33(7)：108－110，112.

[7] 杨飘，董文永. 基于 BERT 嵌入的中文命名实体识别方法 [J]. 计算机工程，2020，46(4)：40－45，52.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. ArXiv, 2017.

[9] Jin Z, Cao J, Zhang Y, et al. News verification by exploiting conflicting social viewpoints in microblogs [C]. Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, 2016: 2972－2978.

[10] Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for microblogs news verification [J]. IEEE Transactions on Multimedia, 2017, 19(3)：598－608.

[11] Jin Z, Cao J, Han G, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C]. ACM International Conference on Multimedia. Mountain View California, 2017: 795－816.

[12] 孙建辉，方向忠. 卷积神经网络的混合精度量化技术研究 [J]. 信息技术，2020，44(6)：66－69.

[13] 丁世飞，齐丙娟，谭红艳. 支持向量机理论与算法研究综述 [J]. 电子科技大学学报，2011，40(1)：2－10.

（责任编辑：丁晓清）

（上接第 136 页）

[5] 付婷，蔡宇翔，李宏发，等. 智能电网中非结构化数据可视化技术研究 [J]. 电网与清洁能源，2019，35(1)：44－48，61.

[6] 江楠，马江涛，郑远攀. 基于 BS 架构的收割机轨迹数据可视化分析系统 [J]. 农机化研究，2020，42(12)：188－193，199.

[7] 王勇，王松，张红英. 基于 B/S 构架的网络结构可视化系统设计与实现 [J]. 计算机工程与应用，2020，56(11)：230－237.

[8] 田静，白珂. BIM 在智能电网资产可视化管理中的应用 [J]. 建筑技术，2019，50(7)：822－825.

[9] 李文芳，程鑫，路强. 一种基于图的电力数据可视分析方法 [J]. 图学学报，2019，40(1)：124－130.

[10] 路强，程鑫，王萍，等. 园区电力数据可视分析系统 [J]. 合肥工业大学学报：自然科学版，2019，42(5)：638－645.

[11] 吴经锋，任双赞，侯喆，等. 基于自适应算法的电力物资抽检系统研究 [J]. 高电压技术，2018，44(2)：66－69.

[12] 王湔璋. 电网工程物资财务管理现状及完善 [J]. 财务与会计，2020，605(5)：73－75.

[13] 陈积光，周蜜. 基于泛在电力物联网的智慧供应链研究 [J]. 控制工程，2020，27(6)：1098－1102.

[14] 王玲. 浅谈海外工程物资采购与管理的认识和体会 [J]. 经济学，2019，2(3)：14－15.

[15] 侯慧，耿浩，肖祥，等. 基于节点综合权值的电力应急物资调度模型研究 [J]. 电力系统保护与控制，2019，47(8)：165－172.

（责任编辑：陈文艳）