如何识别欺骗性垃圾邮件的分析研究

李佩芸[1] 龙法宁[1] 冯伊璐[1] 管方莹[1]
LI Peiyun LONG Faning FENG Yilu GUAN Fangying

摘 要 近年来，电商平台收集的客户商品反馈邮件中常包含欺骗性内容，严重干扰了电商平台的商品质量监管工作。为解决这一问题，本文比较了三种针对欺骗性意见垃圾邮件的特征提取方法，并构建了一个包含负面情绪评论的欺骗性意见垃圾邮件数据集。基于该数据集的实验结果表明，RoBERTa 语言模型能够提取更深层次的文本特征，结合常见的文本分类技术，即可有效检测负面的欺骗性意见。RoBERTa 模型展现了最佳的整体性能，对欺骗性意见垃圾邮件的过滤具有重要价值。

关键词 欺骗性垃圾邮件；垃圾邮件过滤；RoBERTa 语言模型

doi：10.3969/j.issn.1672-9528.2024.09.023

0 引言

消费者越来越多地在网上评价、评论和研究产品。因此，包含消费者评论的网站正成为意见垃圾邮件的目标[1]。目前电商平台销售的产品普遍存在信息不够全面、真实、准确等问题，通过邮件反馈产品信息是建立消费网络平台产品质量管控机制的方式之一。网站收到客户对产品质量评价的邮件往往包括正面意见和负面意见。负面意见较多的产品可以考虑下架。但是，通过邮件反馈商品使用情况获得金钱收益有越来越多的可能性——不适当或带欺诈性评论意见的垃圾邮件，甚至是故意为评价不佳的产品撰写正面评价意见，从而误导电商平台[2]。

虽然其他类型的垃圾邮件检测系统已经应用非常广泛，但遗憾的是，关于欺骗性意见的垃圾邮件检测工作很少。此外，在该领域大多数研究工作专注于检测易于被人类读者识别的内容，例如广告、问题和其他不相关或无意见的文本。本文论述一种潜在危险的欺骗性意见垃圾邮件的识别方法，主要用于判断邮件属于正面意见还是负面意见。如图 1 所示，和其他垃圾邮件检测系统一样，都可看作是短文本的二分类问题，文本分类通过邮件的内容，根据分类算法将邮件归为负面邮件还是正面邮件。通过分类模型提取邮件特征可以发现样本数据特征存在的异同，从而利用分类算法进行正确识别。

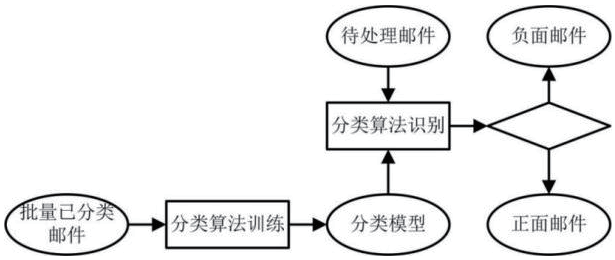

图 1 处理垃圾邮件流程

1 相关工作

目前许多研究人员已经在垃圾邮件检测方面提供了很多方法。这些方法一般包括机器学习、深度学习和一些基于统计学的方法。机器学习和深度学习是解决图像和语言处理等现实问题的分类方法。机器学习方法在少量数据上表现良好，而深度学习方法需要大量数据才能超越机器学习方法的性能[3]。

在传统的机器学习方面，陈亮等人[4]展示了基于 KNN 的方法实现，与前馈神经网络相比具有更高的准确性。姚严志等人[5]结合经典的 TF-IDF 算法，提出了基于类信息的改进的 TF-IDF-CI 算法。Delgado 等人[6]使用欺骗检测辅助神经网络、随机森林的训练学习，为新的研究方向铺平了道路。

在深度学习方面，Faris 等人[7]提出在 Spam Assassin 数据集上，使用前馈神经网络 Krill Herd 算法用于特征提取。结果表明，用于特征提取的 Krill Herd 算法与其他流行的神经网络算法相比具有更好表现。BERT 使用迁移学习，其架构基于 Transformer 模型[8]，迁移学习指的是先训练一个通用任务的模型，然后利用该模型在通用任务

1. 玉林师范学院计算机科学与工程学院 广西玉林 537000

中信息技术与信息化计算机应用获得的知识，通过微调BERT模型将其应用于新任务，例如用于文本信息的特征提取。对BERT的预训练进行了仔细的评估，包括超参数和训练集大小的配置，王乾等人[9]发现BERT其实没有很充分地训练，从而提出了更好地训练BERT的方法，称为RoBERTa，它超过了之前所有已发布的基于BERT的特征提取方法。

2 研究方法

Transformer为自然语言理解和自然语言生成提供通用架构，包括BERT、GPT-2、RoBERTa、XLM、DistilBert、XLNet在内的100多种语言预训练模型，同时支持TensorFlow 2.0和PyTorch机器学习库，可以用于特征提取。把邮件当成一个字符串处理，然后在字符串处理过程中，过滤掉空白符，例如回车符、换行符等，最后遍历全部邮件文件，在词袋模型的基础上，根据邮件内容生产的词汇表对原有句子按照单词逐个进行编码获得新的数据集。如图2所示，使用Transformer模型导入预训练模型，输入编码后的数据集通过编码层来进行特征的提取。

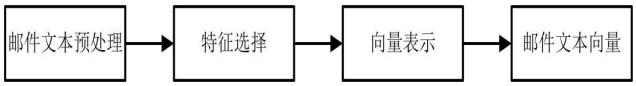

图2 邮件文本的向量表示过程

2.1 特征模型选择

BERT模型：本文实验使用的模型是Transformer通用架构的预训练模型bert-base-uncased，该模型是近年在自然语言领域最具突破性的一项技术，训练分为预训练阶段和微调优化阶段。BERT的第一个预训练任务是Masked LM（MLM），是指在语料库中随机掩盖15%的词汇用于预测任务，这15%的词汇中采用特殊标记替换的概率为80%，随机词汇替换的概率为10%，剩余10%的概率则不进行替换。通过这种掩盖方式，可以在一定程度上避免预训练阶段和微调优化阶段不匹配的问题。模型最大序列长度为40，以容纳最长的可能序列。语料库选择50%的句子对，这部分句子对的第二句就是第一句的下一句，剩余50%的句子对，第二句从语料库中随机抽取。该模型由12个Transformer块组成，每个块都有12个自注意力头和768维的隐藏层。该系统使用Keras函数式API构建和微调。输入层考虑了序列的最大长度，对于每次迭代对原始数据随机舍弃10%，以减少过度拟合。全连接层的两个神经元来表示数据标签中的类别数。因为数据集中的实例数量非常小，所以该模型训练时间短并提供了良好的性能。

RoBERTa模型：本文实验使用了Transformer通用架构的预训练模型RoBERTa-base，该模型的架构与BERT相同，采用深层双向Transformer架构充分获取输入文本语法和语义信息，根据上下文语境不同，生成动态字向量。通过修改扩展BERT模型超参数，舍弃预测下一个句子任务，两个句子通过BERT模型拼接为一个句子对。训练过程中采用更大的预训练数据、预训练步数和批次，提升模型泛化能力。输入文本先分配输入分词器编码成Tokens令牌。然后使用Tokens令牌收集数据特征并进行句子配对，因此能够预测未注释的语言实例中有意隐藏的内容。

GloVe模型：该模型是基于SVD的LSA算法和word2vec算法，并将这两种算法提取的特征合并。该模型认为，语料库中单词出现的统计信息是无监督学习词向量表示的重要依据，既使用了语料库的全局统计特征，也使用了局部的上下文特征。与word2vec模型只考虑上下文相比，GloVe模型会考虑全局特征，效果更好。但是该模型和传统Word2Vec一样，预训练模型所得字向量和字符为一一对应，同一字符无法根据不同语境生成不同字向量。

2.2 实验过程和结果

本文使用了Ott M[10]构建的垃圾邮件数据集，分别包含正面评价和负面评价信息各400条，选择的1600条邮件信息都是由不同用户撰写的。实验时从各数据集中抽取80%的样本作为训练数据，20%的样本作为测试数据。

由于数据集中的消息文本长度不同，需要使所有消息具有相同的长度。如果使用最大邮件的文本长度来填充其他比较短的消息，就会使训练变慢。因此，需要查看训练集中序列长度的分布，以找到合适的填充长度。如图3所示，可以清楚地看到，大部分消息的长度都在300字以内。而最大长度是800。如果选择800作为填充长度，那么所有输入序列的长度都将是800，并且这些序列中的大多数标记将是填充标记，这不会帮助模型学习任何有用的东西。因此，设置300作为填充长度。如果文本的长度大于长度300，则用0填充。

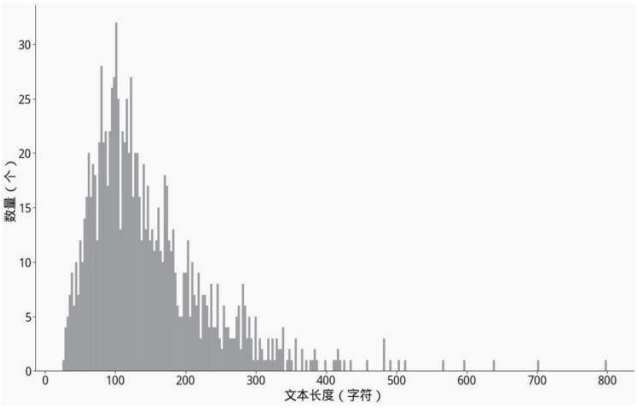

图3 数据集文本长度分布图

本文使用的神经网络模型算法是基于 Pytorch 机器学习框架完成的，研究的课题属于文本分类问题，文本分类问题最常用的评价指标包括准确率、精确率、召回率以及 F_1 值，利用三种特征模型提取特征值，最终通过全连接网络层连接 Softmax 层实现了对电子邮件的分类。

为了验证不同特征提取模型的分类性能，在相同实验环境下进行对比。实验选择自然语言任务模型 BERT 和 RoBERTa 作为预训练模型，这两个模型都共有 12 层，激活函数为 ReLU，迭代次数 epoch 为 20，学习率为 1e-4，数据批次 batch_size 为 32，隐藏层的维度为 768；兼顾全局信息和局部信息的词向量模型 GloVe 嵌入维度为 50，droupout_rate 为 0.3，采用 LSTM 作为神经网络层。为了防止过拟合，本文采用了基于 L2 正则化的交叉熵损失函数来微调模型的训练过程，以追求更好的预测效果。此外，本文还将语境 Transformer 编码器的自注意力权重大小进行归一化处理，同时限定为非负值，控制在 0 和 1 之间，即只允许它对模型的每个位置上的隐藏向量进行正加权组合。

根据表 1，预训练模型的得分都很高，RoBERTa 更是拿到了数据集的最高正确率和 F_1 值。RoBERTa 和 BERT 是基于 Transformer 结构的模型，相比 GloVe，能更加精准地提取隐喻文本中的特征信息。但 BERT 和 RoBERTa 表现出在训练速度上的劣势。如果不考虑模型训练时间，BERT 和 RoBERTa 在处理欺骗性垃圾邮件的分类中就是一个比较好的选择，效果要优于仅使用 LSTM 的 GloVe 模型。

表 1 模型实验结果

模块	正确率 /%	精确率 /%	召回率 /%	F_1 值 /%
GloVe	81.56	93.49	93.89	81.78
BERT	92.62	93.48	93.90	94.56
RoBERTa	99.95	98.23	97.88	98.35

3 结语

本文研究三种提取垃圾邮件特征向量的神经网络模型，经验证，RoBERTa 模型能够更好地提取带欺骗性的垃圾邮件的文本特征，因为 RoBERTa 具有特有的 Tokens 令牌方法和识别词汇表以外单词的能力，几乎不需要对数据集的文本进行预处理（清理）。通过深度学习神经网络，运用预训练模型来自动学习数据特征的方式，比人工手动提取特征更高效，有效地保证数据特征的质量。因此，通过预训练模型编码层的提取特征将是现代垃圾邮件过滤器的核心主流研究方向。未来的研究方向包括为本文提出的方法扩展更多评估指标，采用更多包含负面意见和正面意见的电商产品数据集进行实验，并将其应用到生产环境中。

参考文献：

[1]BOND C F, DEPAULO B M.Accuracy of deception judgments[J].Personality and social psychology review, 2006, 10(3): 214-234.

[2]BULLER D B, BURGOON J K.Interpersonal deception theory[J].Commun theory,1996,6(3):203-242.

[3] 张建 , 严珂 , 马祥 . 基于神经网络的复杂垃圾信息过滤算法分析 [J]. 计算机应用 ,2022,42(3):770-777.

[4] 陈亮 , 朱元凯 , 李长英 . 基于 HHO-KNN 优化算法的垃圾邮件检测研究 [J]. 电脑与电信 ,2022(9):73-77.

[5] 姚严志 , 李建良 . 基于类信息的 TF-IDF 权重分析与改进 [J]. 计算机系统应用 ,2021,30(9):237-241.

[6]DELGADO A A C, GLISSON W, SHASHIDHAR N, et al.Deception detection using machine learning [C]//Proceedings of the Annual Hawaii International Conference on System Sciences. [S.l.]:[s.n.],2021:7122-7130.

[7]FARIS H, ALJARAH I, ALQATAWNA J.Optimizing feedforward neural networks using krill herd algorithm for e-mail spam detection[C]//2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). Piscataway:IEEE,2015:1-5.

[8] 彭毅 , 姜昕宇 . 基于 BERT-DPCNN 文本分类算法的垃圾邮件过滤系统 [J]. 电脑知识与技术 ,2022,18(22):66-69.

[9] 王乾 , 曾诚 , 何鹏 , 等 . 基于 RoBERTa-RCNN 和注意力池化的新闻主题文本分类 [J]. 郑州大学学报 (理学版), 2023 (22): 1-8.

[10]OTT M, CARDIE C, HANCOCK J T.Negative deceptive opinion spam[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.Stroudsburg, PA: Association for Computational Linguistics, 2013:497-501.

【作者简介】

李佩芸（1997—），女，广西玉林人，硕士，研究方向：博弈论与机制设计、机器学习。

龙法宁（1978—），通信作者（email: longfaning@163.com），男，广西玉林人，硕士，高级工程师，研究方向：生物信息、机器学习。

（收稿日期：2024-06-05）