ISSN:2317-0840

文档向量化方法之间的比较:文本数据的案例研究

杰西卡·库布鲁斯利1 †、加布里埃尔·GL·瓦莱诺蒂2 ,

1弗鲁米嫩塞联邦大学统计系、数学与统计研究所。
[2] 弗鲁米嫩塞联邦大学计算机学院。

摘要:近几十年来数字信息的爆炸式增长带来了海量的文本数据。从海量数据中提取知识的兴趣催生了文本挖掘。该领域的挑战之一是将文本语料库转换为数字数据库。这个过程称为文档矢量化,对于自动化信息提取至关重要。这项工作的目标是比较四种文档矢量化方法用于分类目的时的性能。

比较的向量化方法是 Bag of Words (BoW)、TF-IDF 以及 doc2vec、CBOW 和 Skip-gram 两种不同的架构。应用的分类方法有逻辑回归、决策树、随机森林、XGBoost 和感知器。使用的数据集是公开的女性电子商务服装评论数据集,该数据集由 10 个属性组成,本工作考虑了其中三个属性:商品评论文本、评论标题和指示客户是否推荐该产品的分类变量或不。选择了 8,000 份文档的平衡随机样本,其中 4,000 份文档有正面推荐,4,000 份有负面推荐。该数据集分为训练集 (70%) 和测试集 (30%)。性能比较指标是 ROC 曲线下的面积 (AUC)。在比较文档向量化方法时,doc2vec 的两种架构在所有测试的分类方法中都优于其他向量化方法。

关键词:文本挖掘; Doc2Vec; TF-IDF;分类方法。

比较文档的方法:um estudo de caso com bados textuais

摘要:近年来,由于大量数据以文本形式出现,数字化信息不断涌现。我对来自 Min-era ¸ c ao de Texto 的大量数据额外感兴趣。这两个愿望都寄托在“区域”内,并将文本银行转换为数字数据库。此过程将被执行文件扫描,这是自动提取信息的关键。这项工作的目的是比较在分类结束时使用的四种文档扫描方法。向量化的方法有以下几种:BoW、TF-IDF 以及两种不同的架构:doc2vec、CBOW 和 skip。方法论

分类应用程序论坛:回归 Log´ stica、Arvore de Classifica ¸ co、Floresta Aleat´oria、XGBoost 和 Perceptron。公共基础女装电子商务服装评论的基础,包含 10 个属性,其中 3 个考虑因素包括:o 文本 avalia ¸ cao do item,ot´ tulo da avalia ¸,我们可以根据客户推荐或产品类别进行分类。我们共收集了 8,000 份文件,其中 4,000 份文件为积极推荐,4,000 份为消极推荐,分类和区分为三项 (70%) 和一项测试 (30%)。与 ROC 曲线面积 (AUC) 比较的药物。全面比较文档的矢量化方法,作为 doc2vec 的双重设计,呈现出比测试的分类方法更优越的结果。

单词:Minera ¸ c ao de Texto; Doc2Vec; TF-IDF;分类方法。

†作者通讯员: jessicakubrusly@id.uff.br 。

介绍

确实,大多数互联网数据都是非结构化形式,主要是文本。这些数据源自 Facebook 和 Twitter 等社交网络,甚至是投诉和民意调查等公司数据。一个真实的例子是女装电子商务评论1,它由真实客户撰写的评论组成。文本挖掘是数据挖掘的一个分支,它的出现是为了从大量文本数据中处理和提取信息。

情感分析包括对观点、情绪和文本主观性的计算处理（MEDHAT;HASSAN;KORASHY,2014）。其技术大致可分为两类,第一类是基于词典的方法,将单词或表达与积极、消极或中性的感觉联系起来,第二类是机器学习方法。

关于机器学习方法,文档必须以矢量形式表示,以便可以将其用作机器学习方法的输入数据。最直观的替代方法是处理每个文档中的术语（单词）存在或频率（BREIMAN,2001）。但是,还有其他文档矢量化方法,将在本文中讨论。

在此背景下,本研究分析了女装电子商务评论数据库,该数据库包含消费者对特定服装的评论以及表明消费者是否表示该服装的标签。本研究旨在根据客户的文本报告自动预测客户推荐,其目标是比较用作分类方法输入数据的不同文档矢量化方法的质量。

本文的结构如下。第 2 节引用了一些相关研究,第 3 节描述了主要的文本挖掘过程步骤,第 4 节介绍了应用的分类方法,即 Logistic 回归、分类树、随机森林、XGBoost 和 Perpeptron。第 5 节包含分类方法的质量测量,第 6 节描述了数据库,第 7 节介绍了一些有趣的数值结果和分析,最后,第 8 节报告了结论。

相关工作

2022 年,库布鲁斯利等人。 （KUBRUSLY;NEVES;MARQUES,2022)使用相同的数据集,并且仅使用 BoW 向量化方法。该研究的主要目的是比较不同的基于树的分类方法在根据客户推荐对文本进行分类时的性能。结果表明,随机森林和 XGBoost 表现出过度拟合,分类树能够熟练地检测负面评论,但难以处理正面评论,梯度提升显示出稳定的值,测试数据集的 F1 测量值高于 77%。相比之下,本文的重点是比较矢量化方法,而不是分类方法。

2008 年,Sch¨utze、Manning 和 Raghavan（2008)详细讨论了 TF-IDF 和 BoW 模型,为传统文档向量化技术奠定了坚实的基础。这些方法仍然具有现实意义,并经常被用作比较研究的基线。五年后,Mikolov 等人（2013a)提出了 Word2Vec,这是一种生成词向量表示（词的向量表示)的方法。它为后续大量关于词向量表示及其在文本分析中的应用的研究铺平了道路。2014 年,Le 和 Mikolov（2014)提出了一种学习句子和文本文档的向量表示的无监督算法,Kim（2014)展示了卷积神经网络 (CNN) 在句子和文档分类任务中的应用。它展示了深度学习模型在文档向量化中的潜力。

Qasem e Sajid (2022) 对假新闻检测工具进行了研究。他们调查并比较了词袋模型 (BoW) 和词频-逆文档频率模型 (TF-

IDF)方法,使用 N 元语法和三个传统的机器分类器:支持向量机 (SVM)、朴素贝叶斯 (NB) 和决策树 (DT)。结果表明,传统模型仍然是很好的候选模型,并且使用二元模型结合 BOW 和 DT 分类器表现最好,准确率达到 99.74%。

Ling e Chen (2023) 使用来自人类和机器人的推文进行了一项研究。
兴趣在于比较不同词嵌入方法和分类模型的性能。他们使用 f1 分数和混淆矩阵进行评估。结果表明,基于 Transformer 的矢量化方法(包括 Doc2Vec、BERT 和 fastText)在处理不平衡数据时具有强大的功能。

在 Joseph e Yerima (2022) 的研究中,通过评估常用的词嵌入技术在公开的正常短信和垃圾邮件数据集上的性能,对用于检测垃圾邮件的流行词嵌入技术进行了比较分析。使用 5 种不同的机器学习分类器(即多项式朴素贝叶斯 (MNB)、KNN、SVM、随机森林和额外树)研究了词嵌入技术的性能。根据研究中使用的数据集,N-gram、BOW 和过采样的 TF-IDF 获得了最高的 F1 分数,正常短信为 0.99,垃圾邮件为 0.94。

文本挖掘

文本挖掘基本上是从非结构化文本文档中提取重要模式或知识的过程。这个过程可以分为两个主要步骤: 细化,将原始文本数据库转换为数字数据库;信息提取过程,包括使用传统统计工具从精炼数据库中检测模式(FRITSCH;GUENTHER;WRIGHT,2019)。

文本预处理

文本数据库的主要细化步骤,也称为预处理技术,是:停用词删除;正常化。下面对每一项进行简要说明。

标记化是第一个预处理阶段,旨在从中提取最小文本单元
自由文本。这些单位称为标记,通常指单个单词。
停用词是语言中最常用的词。它们没有语义价值,仅有助于对文本的一般理解。停用词通常以冠词、介词、标点符号、连词和代词为特征。通常使用预先建立的列表,称为停用词表。删除停用词可大大减少标记数量并改善要执行的分析。

规范化是将具有相同模式的单词分组的过程。主要的标准化方法是词干提取和词形还原,对这些术语的进一步解释可以在 Goodfellow, Bengio e Courville (2016) 中找到。这里将应用词形还原方法,例如,替换术语“计算”的标记“计算”、“计算”和“计算”。此过程如图 1 所示。

词袋

按照这些步骤,每个文档随后被转换成一个词袋 (Bag of Words - BoW)。考虑由 n 个文档组成的文本数据库,这些文档总共包含 m 个词。为了降低此表示的维度,需要执行词条选择。Hastie 等人 (2009) 提出的词条选择建立了一组重要的数据库词条。不重要的词条具有较低的语义价值,在文档集中出现的频率非常高或非常低,因此在分析中不予考虑。

图 1:文本预处理示例。

资料来源:作者。

在选择完术语后,考虑一个由 n 个文档组成的文本数据库,这些文档总共包含 p 个术语。n × p 矩阵 A 称为文档术语矩阵,其中每个元素ai,j表示术语 j 在文档 i 中出现的频率。该矩阵的每一行都是文档的向量表示。每列对应一个术语,可以理解为文档属性。因此,该方法中文档的向量表示将由以下公式给出:

$$d^{1}_{*} = (ai,1, ai,2, \cdots, ai,p) \quad (1)$$

图 2:词袋向量表示示例。

	love	dress	way	style	disappoint	quality	color	poor	soft	...
1	4	1	1	1	0	0	0	0	0	
2	1	1	0	1	1	2	1	1	0	
3	0	0	0	1	0	0	1	0	1	
4	1	2	0	0	0	0	0	0	0	
⋮										nxp

来源:作者。

TF-IDF

TF-IDF,代表词频-逆文档频率,是自然语言处理和信息检索中广泛使用的技术。它是量化文档集合中术语（单词或短语)重要性的强大工具。 TF-IDF 对于文档检索、文本分类和信息检索等文本分析任务特别有价值。

TF（词频)衡量文档中词的频率。T Fi,j表示词 j 在文档 i 中出现的频率相对于文档 i 中词的总数量

文档 i。本质上,它量化了术语与特定文档的相关性。

$$TFi, j = i \text{ 中的} \frac{j \text{ 出现在 i 中的次数}}{\text{项数}} = \frac{,j}{\sum_{j=1}^{p} ai,j}$$

另一方面,IDF（逆文档频率)评估术语在文档集合中的重要性。 IDFj计算包含术语 j 的文档分数的对数倒数。许多文档中常见的术语会获得较低的 IDF 分数,而罕见或独特的术语会获得较高的分数。

$$IDFj = \ln \frac{\text{文件总数}}{\text{包含术语 j 的文档数量}} = \ln \frac{n}{\sum_{=1} IN \ (ai,j)}$$

其中N是正整数集合,I 是指示函数。

通过结合 TF 和 IDF 组件,可以计算文档中某个术语的 TF-IDF 分数。此分数突出显示在文档中频繁出现且在整个文档集合中具有独特性的术语。在实践中,TF-IDF 有助于识别区分文档的关键字或相关术语,使其成为各种基于文本的应用程序的必备技术。

$$TF\text{-}IDFi,j = \frac{T\ Fi,j}{\text{以色列国防军}}$$

因此,该方法将提供以下文档向量表示。

$$^{2} = (TF\text{-}IDFi,1,\ TF\text{-}IDFi,2,\ \ldots,\ TF\text{-}IDFi,p) \quad (2)$$

Doc2Vec

Doc2Vec 是自然语言处理 (NLP) 中一种流行的文档向量化技术。它是 Word2Vec 的扩展,旨在捕获整个文档的语义含义,使其成为各种 NLP 任务的宝贵工具。与以前的方法不同,doc2vec 使用所有 m 个术语的集合作为输入。在这种方法中,无需执行术语选择即可将基数从 m 个术语减少到 p 个术语,因为降维是通过神经网络隐藏层中的神经元数量来执行的。

Word2Vec 由 Tomas Mikolov 及其团队开发（MIKOLOV 等,2013a）,彻底改变了计算机理解和表示大型文本语料库中的单词的方式。它的工作原理是具有相似含义的单词通常出现在相似的上下文中。这项工作中使用的两种架构是连续词袋 (CBOW) 和 Skip-gram（MIKOLOV 等,2013b）。

CBOW 架构根据上下文单词预测目标单词,它通过考虑单词周围的单词来学习理解单词,如图 3 所示。

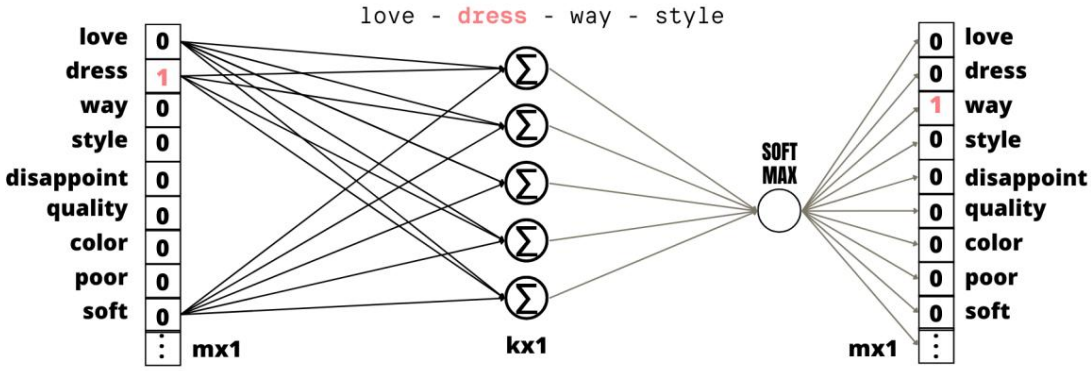
另一方面,Skip-gram 架构基于上下文单词进行预测,它捕获使用单词的上下文,如图 4 所示。在这两种架构中,每个单词都由与相应单词关联的突触权重值表示。

在每个单词已经有了 word2vec 提供的向量表示（图 5 中红线表示)之后,现在可以训练 doc2vec 网络来为文档建立向量表示。Doc2Vec 为语料库中的每个文档分配一个唯一的向量表示,它在训练过程中将单词向量与文档向量相结合,确保保留文档中单词的上下文（图 5）。

文档向量表示也将由突触权重给出,但在这种情况下,
那些与文件相关的

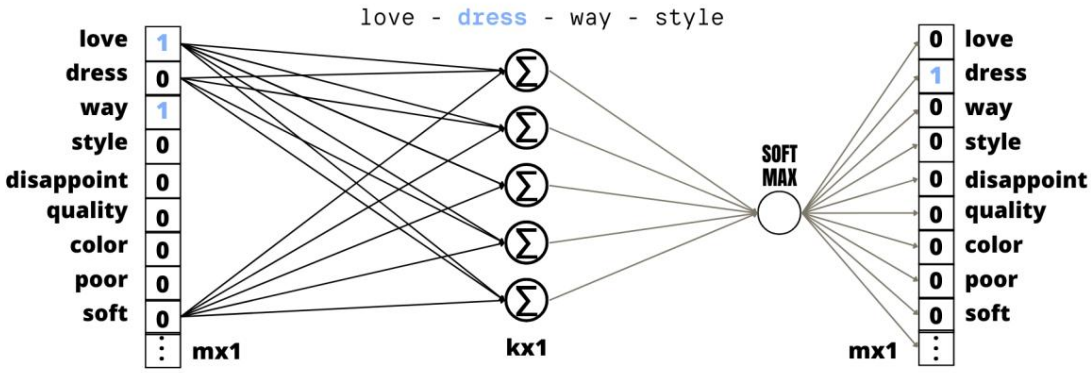
$$d^{3}_{\text{食}} \text{wc wc} = \text{w i,1 , i,2 , . 。 。 ,我知道} \qquad \text{和 d} \quad ^{4}_{\text{食}} \text{ws ws} = \text{wi ,1 , i,2 , } \cdots, \text{i,k}$$

图 3:CBOW 架构示例。



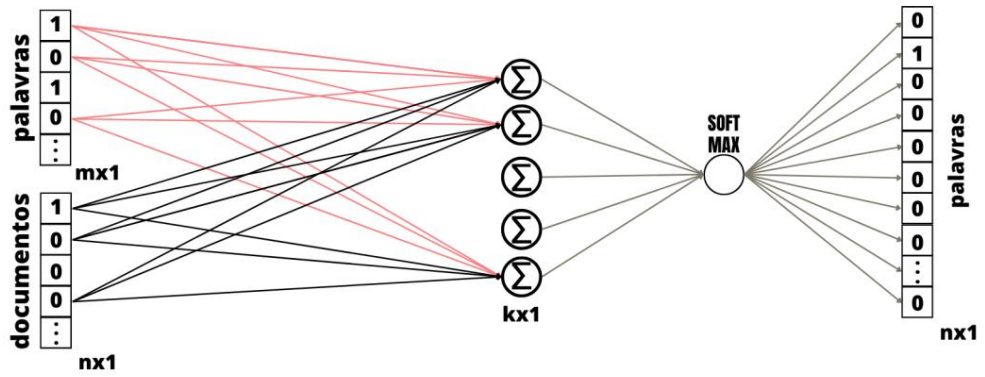
资料来源:作者。

图 4:Skip-gram 架构示例。



资料来源:作者。

图 5:doc2vec 架构示例。



来源:作者。

其中 k 是唯一隐藏层中神经元的数量,w 和 w分别是考虑 CBOW 和 Skip 架构的从文档 i 到神经元 l 的突触权重$_{i,l}^{C}$,$_{i,l}^{s}$。

为了更公平地进行比较,p（术语选择后的术语数量)和 k（doc2vec 神经网络隐藏层中的神经元数量)的值相等。因此,所有向量表示都将具有相同的维度,从现在开始将称为 p。

分类方法

这项工作的重点是比较不同的文档矢量化方法。为了进行这种比较,使用了五种不同的分类方法,本节将对其进行简要说明。

考虑由 n 个对象（文档)组成的宇宙,这些对象由 p 个属性描述,并且每个对象 i 属于已知类Yi = {0, 1}。分类方法旨在定义一个数学模型,当已知新对象的 p 个属性时,该模型能够预测新对象的类别。实际上,分类方法返回属于类别 1 的概率。

逻辑回归 (MCCULLAGH,2019;HASTIE 等,2009) 是一种广义线性模型,是一种用于对二元结果或分类数据进行建模的基本统计和机器学习技术。与预测连续数值的线性回归不同,当因变量为二元或分类时使用逻辑回归。它模拟事件发生的概率,例如客户是否推荐产品。

由 Breiman 等人提出。 (1984),树模型是一种利用树结构对数据集进行递归划分的分类方法。一旦输入数据被分割,就可以通过每个分区中的简单分类方法进行预测,例如主导类别或参考类别的流行率。

随机森林（BREIMAN,2001)是为了改进分类树模型的预测而创建的装袋分类模型。它由分类树的集合组成,其中每棵树都是根据由 ~n < n 个对象组成的较小数据集构建的。 n~ 个对象是从 Bagging 策略中选择的,例如 Bootstrap Sampling (SUTTON, 2005)。

对于森林中的每棵树,每次分割时,都会从 p 个变量中随机选择 p < p。此分区中仅考虑选定的 p 个变量。生成多棵树后,将这些结果组合起来以提供最终预测。

随机森林方法是一种装袋算法。在这些算法中,树并行生长以获得所有树的平均预测。另一方面,梯度提升采用顺序方法获取预测。在梯度提升方法中,每棵决策树都会预测前一棵决策树的误差（AYYADEVARA,2018 年）。XG-Boost 方法是升级的梯度提升树算法,可以灵活处理稀疏数据和缺失值（LIN,2020 年）。该系统在单台机器上的运行速度比现有的流行解决方案快 10 倍以上,并且在分布式或内存有限的设置中可扩展到数十亿个示例。它还结合了正则化模型以防止过度拟合（CHEN; GUESTRIN,2016 年）。

支持向量机 (SVM) 方法由 Cortes e Vapnik (1995) 开发。它是一种用于分类任务的强大机器学习算法。 SVM 的工作原理是使用核函数找到一个超平面,该超平面可以有效地将属于高维特征空间中不同类的数据点分开。 SVM 可以处理线性和非线性数据,并且在应用适当的正则化时,它可以防止过度拟合。

Finelly,多层感知器,也称为多层神经网络 (MLP)（GOODFELLOW;BENGIO;COURVILLE,2016)是简单感知器的扩展,由多层神经元组成,包括输入层、一个或多个隐藏层和输出层。每一层中的每个神经元都与下一层中的所有神经元相连,使其成为一种深度学习模型。这种方法可以捕捉输入数据中的复杂关系,使其适合更具挑战性的分类和回归任务。值得注意的是,神经网络模型在隐藏单元中实际上是一个线性逻辑回归模型,所有参数都是通过最大似然估计的（HASTIE 等,2009）。

质量措施

受试者工作特征 (ROC) 曲线是用于评估分类方法性能的基本质量度量。它为不同阈值设置下的真阳性率（敏感性)和假阳性率（1-特异性)之间的权衡提供了有价值的见解。 ROC 曲线在评估二元分类模型（例如本研究中使用的模型)时特别有用。

ROC 曲线是通过在不同阈值下绘制 y 轴上的真阳性率 (TPR) 与 x 轴上的假阳性率 (FPR) 来创建的。TPR 表示模型正确识别的真阳性的比例,而 FPR 表示错误地分类为阳性的假阳性的比例。

完美的分类模型的 ROC 曲线应该达到图表的左上角,表明灵敏度高、假阳性率低,曲线下面积 (AUC) 等于 1.0。相反,随机猜测模型的 ROC 曲线非常类似于对角线,AUC 为 0.5。

AUC 代表 ROC 曲线下面积,是 ROC 曲线得出的另一个重要质量指标。它量化了分类模型的整体性能。AUC 值为 1.0 表示分类器完美,而 AUC 值为 0.5 表示模型的表现不比随机概率好。

ROC 图的一个重要点是它们衡量分类器产生良好相对实例分数的能力。分类器不需要产生准确的、经过校准的概率估计;它只需要产生相对准确的分数来区分正面和负面实例（FAWCETT,2006）。

数据集

女装电子商务评论被用作本研究的数据集,并围绕客户撰写的评论进行。该数据集包含 23,486 行和 10 个特征变量。每行对应一个客户评论,并包含以下变量： 服装 ID;年龄;标题;审查文本;评分;推荐的IND;积极反馈计数;部门名称;部门名称;和类名。在数据库的 23,486 行中,19,314 行涉及推荐项目,而其他 4,172 行涉及非推荐项目。

这里只考虑了三个变量:评论文本,评论正文的字符串变量； Title,评论标题的字符串变量;推荐IND,二进制变量,表示客户是否推荐该产品,其中1推荐,0不推荐。变量标题和评论文本被连接起来,以便为分析添加更多丰富的信息。然后,仅使用文本作为分类方法的属性。

首先,将数据集随机分为70%作为训练集,30%作为测试集。由于原始数据库包含比非推荐对象更多的推荐对象,因此不应使用所有文档。为了选择平衡集并尊重 70/30 的比例,训练集中的文档数量为 5,674,测试中的文档数量为 2,445。

结果

这项研究是使用 R 程序（R 核心团队,2019)进行的。 tidytext (SILGE; ROBINSON, 2016)、tm (FEINERER; HORNIK; MEYER, 2008) 和 textstem (RINKER, 2018) 包用于文本预处理、BoW 和 TF-IDF 表示。 word2vec (WIJFFELS, 2021b) 和 doc2vec (WIJFFELS, 2021a) 包用于单词和文档嵌入向量表示。 rpart (THERNEAU; ATKINSON, 2018)、randomForest (LIAW; WIENER, 2002)、xgboost (CHEN et al., 2023) 和 Neuralnet

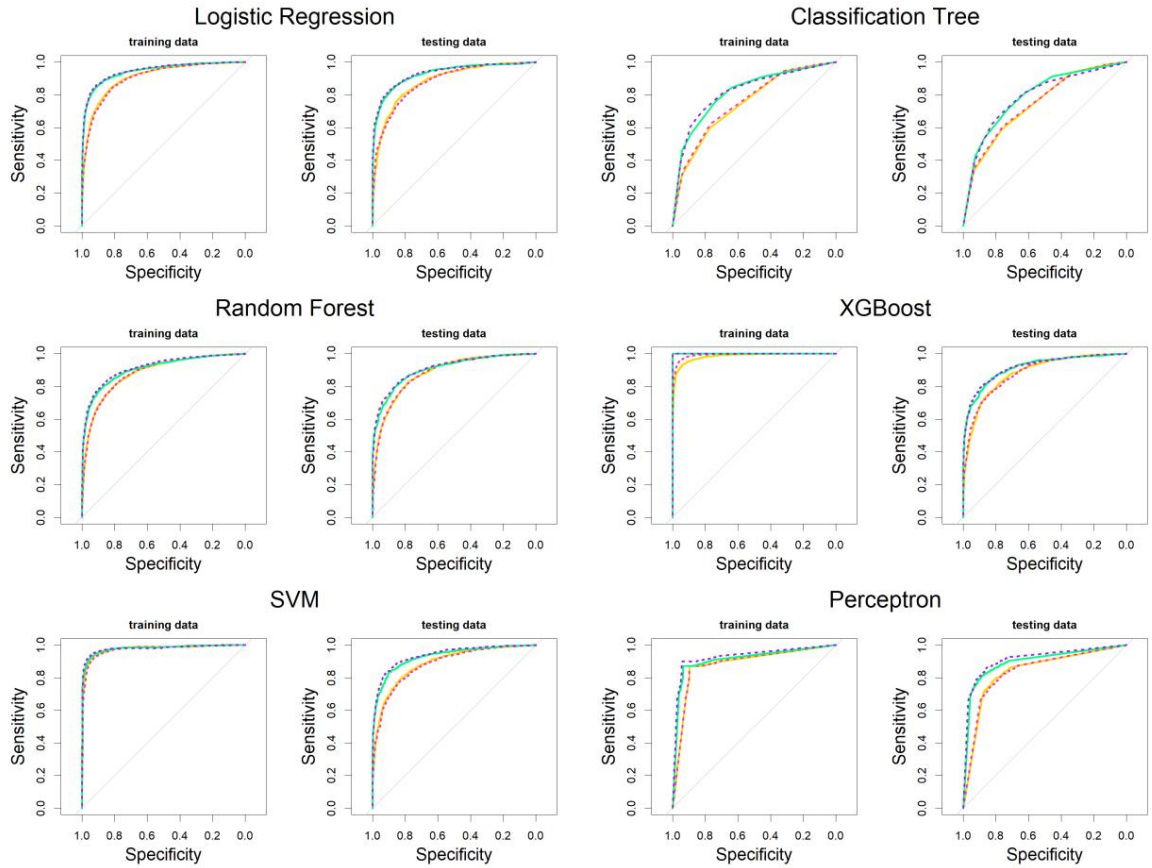
（FRITSCH;GUENTHER;WRIGHT,2019)包分别用于分类树、随机森林、XGBoost 和 Perseptron 分析。pROC（ROBIN 等，2011)包用于 ROC 曲线和 AUC 计算。

文本预处理部分中描述的预处理是针对训练数据集进行的。删除停用词并根据 Mechura 的英语词形还原列表进行规范化过程。在此过程之后，训练文本数据库包含 5,675 个文档和 7,066 个不同术语。

对于词袋 (BoW) 和 TF-IDF 向量表示，执行术语选择过程以选择前 200 个最常见的术语。对于 Doc2Vec、CBOW 和 Skip-gram 两种架构，隐藏层都使用了 200 个神经元。结果是，本案例研究中考虑的所有向量表示的维度均为 200。

所有分类方法均使用默认参数值在 R 中运行。应用的多层感知器模型考虑了具有单个神经元的单个隐藏层。 SVM 方法使用径向核运行。对于四种向量表示中的每一种，总共运行了 6 种分类方法：Logistic 回归、分类树、随机森林、XGBoost、SVM 和感知器。

图 6：分类方法的 ROC 曲线



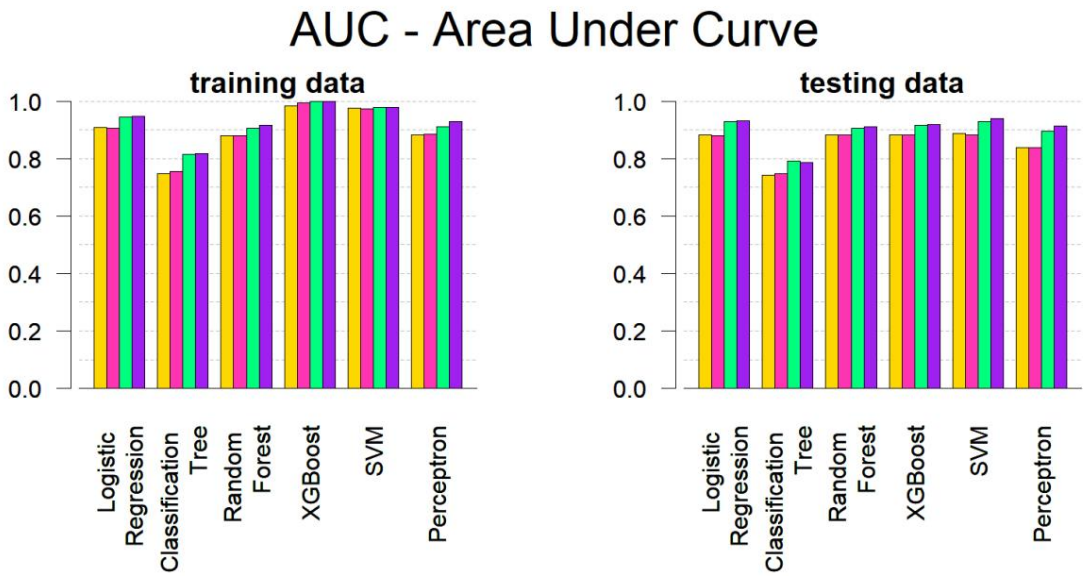
来源：作者。

图 6 显示了所有四种矢量化方法以及六种分类方法中每种方法的 ROC 曲线。 BoW 以黄色实线表示，TF-IDF 以粉色虚线表示，Doc2Vec-CBOW 以绿色实线表示，Doc2Vec-Skip-gram 以紫色虚线表示。

对于本研究中测试的所有分类方法，与 BoW 和 TF-IDF 相比，doc2vec 的两种架构都取得了更好的结果。当输入向量由 BoW 和 TF-IDF 定义时，ROC 曲线还表明分类方法的性能相似。两种 doc2vec 架构的分类方法的性能

也类似。

图 7:分类方法的 AUC 值



资料来源:作者。

图 7 显示了训练和测试数据集的每种分类方法和每种向量化方法的 AUC 值的条形图。这些数字证实了之前提出的结果,即两种 Doc2Vec 架构产生了更好的性能。此外,BoW 和 TF-IDF 提供的矢量化产生相似的结果。

结论

通过消费者评论（即涉及推荐或不推荐产品的自由文本)进行文本挖掘分析。目标是比较当输入数据由不同文档向量化方法生成时分类方法的质量。这项工作中比较的矢量化方法有:Bag of Words (BoW)、TF-IDF,以及两种不同架构的 Doc2Vec、CBOW 和 Skip-gram。

所采用的分类方法包括逻辑回归、决策树、随机森林、XGBoost、SVM 和感知器。所用的数据集是公开的女性电子商务服装评论数据集,该数据集包含 10 个属性,本文考虑了其中三个属性:商品评论文本、评论标题以及表示客户是否推荐该产品的分类变量。选择了 8,000 份文档的平衡随机样本,其中 4,000 份文档有正面推荐,4,000 份文档有负面推荐。该数据集分为训练集（70%)和测试集（30%）。

性能比较指标是 ROC 曲线下面积 (AUC)。

当输入数据由各种向量化方法生成时,分类结果没有显著差异。同样,分类器对向量化方法 CBOW 和 Skip-gram 也产生了类似的结果。在比较所有四种向量化方法时,Doc2Vec 的两种架构（CBOW 和 Skip-gram)在所有测试的分类方法中都优于其他向量化方法。

致谢

谨向弗鲁米嫩塞联邦大学（UFF)表示诚挚的谢意
感谢它使这项工作成为可能,并感谢它对学术卓越的持续承诺。

参考

AYYADEVARA,VK Pro 机器学习算法。Apress:美国加利福尼亚州伯克利,Springer,2018 年。

BREIMAN,L. 随机森林。机器学习,Springer,第 45 卷,第 1 期,第 5-32 页,2001 年。

BREIMAN,L.;FRIEDMAN,J.;STONE,CJ;OLSHEN,RA 分类和回归树。[Sl]:CRC 出版社,1984 年。

CHEN, T.;GUESTRIN, C. Xgboost:可扩展的树提升系统。在:第 22 届 acm sigkdd 国际知识发现和数据挖掘会议论文集。[Sl: sn],2016 年。第 785–794 页。

CHEN, T.; HE, T.; BENESTY, M.; KHOTILOVICH, V.; TANG, Y.; CHO, H.; CHEN, K.; MITCHELL, R.; CANO, I.; ZHOU, T.; LI, M.; XIE, J.; LIN, M.; GENG, Y.; LI, Y.; YUAN, J. xgboost:极端梯度提升。[Sl],2023 年。R 包版本 1.7.3.1。可用： https://CRAN.R-project.org/package=xgboost 。

科尔特斯,C.； VAPNIK,V。支持向量网络。机器学习,Springer,第 20 卷,第 14 页。 273-297,1995。

FAWCETT,T. ROC 分析简介。模式识别快报,Elsevier,第 27 卷,第 8 期,第 861-874 页,2006 年。

FEINERER, I.;HORNIK, K.;MEYER, D. 文本挖掘基础设施。《统计软件杂志》,第 25 卷,第 5 期,第 1-54 页,2008 年 3 月。网址： http://www.jstatsoft.org/v25/i05/ 。

FRITSCH,S.;GUENTHER,F.;WRIGHT,MN Neuralnet:神经网络训练。[Sl],2019 年。R 包版本 1.44.2。发布于： https://CRAN.R-project.org/package=neuralnet 。

古德费洛,I.;本吉奥,Y.； COURVILLE, A. 深度学习。 [Sl]:麻省理工学院出版社,2016。

哈斯蒂,T.;蒂布希拉尼,R.;弗里德曼,JH； FRIEDMAN, JH 统计学习的要素:数据挖掘、推理和预测。 [Sl]:施普林格,2009 年。第 2 卷。

JOSEPH,P.;YERIMA,SY 用于短信垃圾邮件检测的词嵌入技术的比较研究。在:IEEE。2022 年第 14 届计算智能与通信网络国际会议 (CICN)。[Sl],2022 年,第 149-155 页。

KIM,Y. 用于句子分类的卷积神经网络。arXiv 预印本 arXiv:1408.5882,2014 年。

库布鲁斯利,J.;阿拉巴马州内维斯； MARQUES, TL 使用基于树的方法对文本电子商务评论进行统计分析。开放统计杂志,科学研究出版,第 12 卷,n。 3,第 3 页。 357–372,2022。

LE,Q。 MIKOLOV, T. 句子和文档的分布式表示。在:PMLR。
机器学习国际会议。 [Sl],2014 年。 1188–1196。

LIAW,A.;WIENER,M. 通过随机森林进行分类和回归。R News,第 2 卷,第 3 期,第 3 页。
2002 年 18 日至 22 日。发布时间： https://CRAN.R-project.org/doc/Rnews/ 。

林X。基于自然语言处理的电子商务客户评论情感分析。见:2020年第二届大数据与人工智能国际会议论文集。 [Sl:
sn],2020 年。 32-36。

林,J.； CHEN, Y. 在线 Twitter 机器人检测:平衡和不平衡数据的矢量化和分类方法的比较研究。工程档案,2023 年。

MCCULLAGH, P. 广义线性模型。 [Sl]:劳特利奇,2019。

MEDHAT, W.;HASSAN, A.;KORASHY, H. 情感分析算法与应用:调查。《艾因·夏姆斯工程杂志》,Elsevier,第 5 卷,第 4 期,
第 1093-1113 页,2014 年。

米科洛夫,T.;陈,K。科拉多,G.； DEAN, J. 向量空间中单词表示的有效估计。 arXiv 预印本 arXiv:1301.3781,
2013。

米科洛夫,T.;苏茨克弗,I.;陈,K。科拉多,GS； DEAN, J. 单词和短语的分布式表示及其组合性。神经信息处理系统的进
展,第 26 卷,2013 年。

QASEM, AE;SAJID, M. 探索带有 bow 和 tf-idf 表示的 n-gram 对检测假新闻的影响。在:2022 年商业和工业数据分析国际
会议 (ICDABI)。[Sl: sn],2022 年。第 741–746 页。

R 核心团队。R:统计计算的语言和环境。奥地利维也纳,2019 年。地址： https://www.R-project.org/ 。

RINKER,TW textstem:用于对文本进行词干提取和词形还原的工具。纽约州布法罗,2018 年。
版本 0.1.4。请访问： http://github.com/trinker/textstem 。

罗宾,X.;图尔克,N.;海纳德,A.;蒂贝尔蒂,N.;利萨切克,F.;桑切斯,J.-C.； MìuLLER, M. proc:r 和 s+ 的开源包,用于分析
和比较 roc 曲线。
BMC 生物信息学,第 12 卷,第 77 页,2011 年。

舒茨,H.;曼宁,CD;拉加万,P. 信息检索简介。
[Sl]:剑桥大学出版社剑桥,2008 年,第 39 卷。

SILGE,J.;ROBINSON,D. tidytext:使用 r 中的整洁数据原则进行文本挖掘和分析。JOSS,The Open Journal,
第 1 卷,第 3 期,2016 年。发布于： http://dx.doi.org/10.21105/joss.00037 。

SUTTON,CD,分类和回归树、装袋和提升。《统计手册》,Elsevier,第 24 卷,第 303-329 页,2005 年。

THERNEAU,T.;ATKINSON,B. rpart:递归分区和回归树。[Sl],2018 年。R 包版本 4.1-13。发布于： https://CRAN.R-
project.org/package=rpart 。

WIJFFELS, J. doc2vec:句子、文档和主题的分布式表示。 [Sl],2021 年。R 包版本 0.2.0。分发： https://CRAN.R-
project.org/package=doc2vec 。

_____。 word2vec:单词的分布式表示。 [Sl],2021 年。R 包版本 0.3.4。
分发： https://CRAN.R-project.org/package=word2vec 。