融合多特征的产品垃圾评论识别

吴 敏[1],何 珑[2]

(1.福州大学 数学与计算机学院,福建 福州350108;
2.福州大学 信息化建设办公室,福建 福州 350108)

摘 要:针对JINDAL N等人新近提出的利用逻辑回归模型识别产品垃圾评论的检测方法中使用过多产品评论特征这一问题,分析了解决方法,并提出对特征进行显著性检验。通过对亚马逊数据集的实验结果表明,采用显著性特征建立的回归模型优于所有特征建立的模型。新模型不仅解决了上述问题,减少了计算量,而且整体性能不变,这表明以显著性特征建模有助于提高模型的检测质量。

关键词:逻辑回归;产品垃圾评论;显著性检验

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2012)22-0085-03

Fuse multi-features to identify product review spam

Wu Min[1], He Long[2]

(1.College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China;
2.Information Construction Office, Fuzhou University, Fuzhou 350108, China)

Abstract: To solve the problem of overusing features in the product review spam identification based on logistic regression recently proposed by JINDAL N et al., we take significance testing on these features. Our experiments on the Amazon dataset show that the new regression model based on the significant features is better than the model based on the whole features. This new model not only solves the problem mentioned above, but also achieves the same performance with lower calculation cost; it shows that modeling on the significant features contributes to improving the detection quality.

Key words: logistic regression(LR); product review spam; significance testing

近几年,随着互联网的发展,人们越来越喜欢在网络上表达自己的观点。他们可以在购买商品的同时在各大商业网站、论坛以及博客发表评论。这些观点信息对其他潜在用户至关重要。

由于网络的开放性,人们可以在网站上任意书写评论,这导致评论的质量低下,甚至产生垃圾评论,即由一些用户蓄意发表的不切实际、不真实的、有欺骗性质的评论,其目的是为了提升或者诋毁某一产品或某一类产品的声誉,从而误导潜在消费者,或者干扰评论意见挖掘和情感分析系统的分析结果[1]。正面评论可以提高产品销售额,还可以提高公司的名声,负面评论则可以诋毁竞争对手。这就为垃圾评论发表者提供了足够的动机。2007年,JINDAL N和Liu Bing首次对垃圾评论检测进行相关研究[1-2]。

1 相关工作

目前,在线观点的分析已经成为一个热门的研究主题。然而,现有工作主要集中在利用自然语言处理和数据挖掘技术来抽取和总结评论观点[3-4],对评论的特征以及评论者的行为研究较少,而这些却是观点挖掘的必要前提。

目前的研究工作中已经取得了很多成果,JINDAL N等[1-2]将垃圾评论分为三类:欺骗性的评论(Untruthful Opinion);不相关的评论(Reviews on Brands Only);非评论信息(Non-Reviews)。之后,他们收集了评论文本、评论发表者和产品3个方面共36个特征,采用人工标记训练集的方法,应用Logistic回归建立机器学习模型来识别第二类和第三类垃圾评论;对于第一类垃圾评论,则通过识别重复性的评论,将重复性评论作为正向的训练集建立机器学习模型来识别。此方法取得了不错的效果,但使用的特征过多,不仅增加了计算量,而且可能使得模型不够稳定。因此,本文利用重复评论建立模型,提出对特征进行显著性检验,以获取的显著性特

征建立更加稳定的回归模型。实验结果表明,新模型不仅有效地减少了计算量,而且效果优于所有特征建立的模型。

2 融合多特征的产品垃圾评论识别方法

2.1 垃圾评论检测

在JINDAL N和Liu Bing的工作中,他们将垃圾评论分为三类,本文主要致力于检测第一类垃圾评论,即欺骗性的评论(Untruthful Opinion)。

2.1.1 检测重复评论

对于第二和第三类垃圾评论,可以通过评论内容来识别。然而,仅仅通过人工阅读一个评论来判别它是否具有欺骗性是极其困难的,这是由于垃圾评论发表者可以通过仔细伪装使评论看起来和其他正常评论一样。

因此,本文利用JINDAL N等[2]提出的以下3种重复评论(包括近似重复)来检测第一类垃圾评论:(1)不同用户对同一产品发表的重复评论;(2)相同用户对不同产品发表的重复评论;(3)不同用户对不同产品发表的重复评论。

同一用户对同一产品的重复评论,有可能是因为用户多次点击提交造成的,也有可能是用户为了修改之前的评论。为此,只保留同一用户对同一产品的最新评论。

重复和近似重复评论的识别使用的是Shingle Method[5]。首先,对所有评论建立2-Gram语言模型,然后对两个评论A、B计算相似值$J(A,B)$,公式如下:

$$J(A,B)=(A\cap B)/(A\cup B)$$

当两个评论的相似度在90%以上时,把它们当作重复评论。

2.1.2 模型的建立

本文使用R统计软件来建立逻辑回归模型,并将AUC(Area under ROC Curve)作为分类结果的评价指标。AUC是一个用于评价机器学习模型质量的标准指标。

为了建立模型,需要构建训练数据,为此,本文使用了JINDAL N和Liu Bing总结的特征来表示评论,具体特征见参考文献[4]。其中本文对部分特征的处理可能与他们的方法存在差异:对于特征F10、F11,本文中的这些词来自知网(Hownet)提供的最新情感词词典,与JINDAL N等的词典不同;特征F26的评论数少于3不予判断,当做0。

2.2 显著性检验

根据样本得到的Logistic回归模型需要经过检验才能说明影响因素对事件发生的影响是否具有统计学意义。特别是当影响因素比较多时,需挑选出与事件发生确实有关或关系更密切的影响因素,以建立更加稳定的回归模型。

对逻辑回归模型:

$$\ln\left[\frac{P}{1-P}\right]=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n$$

检验模型中自变量X_j是否与反应变量显著相关。

假设检验:

$$H_0:\beta_j=0 \qquad H_1:\beta_j\neq 0$$

拒绝H_0则表示事件发生的可能性依赖于X_j的变化。

假设检验主要有Wald检验(Wald Test)、似然比检验(Likelihood Ratio Test)和计分检验(Score Test)等方法,对回归系数进行显著性检验时,通常使用Wald检验,可通过将统计量取平方得到,其公式为:

$$\text{Wald}=(\beta_j/se_{\beta_j})^2$$

这个单变量Wald统计量服从自由度等于1的χ^2分布。其中,Z统计量其实就是某个自变量所对应的回归系数与其标准误差的比,即:

$$Z=\beta_j/se_{\beta_j}$$

在大样本情况下,可以直接用Z检验来作为个别自变量参数估计的统计检验。在本文中,通过Wald检验的方法计算特征的显著性。

3 实验结果及其分析

3.1 语料

本文使用了由Liu Bing在网上提供的来自Amazon的数据,如表1所示。使用该数据集的原因是它的数据量大,且Amazon作为最成功的商业网站之一,拥有着相当长的历史,使用该网站提供的数据集是相当合理的。本文的实验就是从mproducts领域中随机抽取了部分评论进行研究,在保留同一用户对同一产品的最新评论后,获得7 315个评论,其中重复评论有128个。

表1 Amazon数据集

类别	评论	产品	评论数	产品总计
总计	5 838 032	1 195 133	2 146 048	6 272 502
书籍	2 493 087	637 120	1 076 746	1 185 467
音乐	1 327 456	221 432	503 884	888 327
DVD/VHS	633 678	60 292	250 693	157 245
mProducts	228 422	36 692	165 608	901 913
自统计数据	7 315	4 865	5 032	–

Amazon网站的每个评论由8部分组成:Product ID;Reviewer ID、Rating、Date、Review Title、Review Body、Number of Helpful Feedbacks和Number of Feedbacks。

3.2 显著性特征抽取

本文利用R软件对上文提到的36个特征逐一进行Wald检验,以筛选出有显著意义的特征,筛选部分结果如表2所示。其中,Intercept为回归模型的截距;Estimate代表每个特征对应的系数值β_j;Std. Error为对应特征的标准误差se_{β_j};z value为Z统计量β_j/se_{β_j};Pr(>|z|)即为接受原假设$\beta_j=0$的概率。显著性规则:0<'***'<0.001<'**'<0.01<'*'<0.05<'.'<0.1<' '<1,即'***'特征最显著,之后逐级递减。

以上显著性特征中,'***'和'**'显著性特征主要为F13~F16以及F24~F30,即文本特征和评论者特征,这说明评论文本内容和评论者行为在识别重复评论中发

表 2 显著性检验结果

Coefficients	Estimate	Std.Error	z value	Pr(>|z|)	显著性
Intercept	-4.866e+00	1.035e+00	-4.702	2.58e-06	***
F5	4.678e-04	7.547e-05	6.198	5.71e-10	***
F13	4.270e+01	8.448e+00	5.054	4.33e-07	***
F14	-5.027e+01	1.514e+01	-3.321	0.000896	***
F15	1.597e+01	5.362e+00	2.978	0.002905	**
F16	-1.528e+01	5.338e+00	-2.862	0.004209	**
F19	-8.825e-01	4.079e-01	-2.164	0.030482	*
F24	-5.983e-01	1.367e-01	-4.379	1.19e-05	***
F25	-1.034e+00	3.421e-01	-3.023	0.002502	**
F26	2.870e+00	4.689e-01	6.121	9.32e-10	***
F27	4.209e+00	6.854e-01	6.141	8.19e-10	***
F28	2.702e+00	6.875e-01	3.931	8.47e-05	***
F29	3.133e+00	1.131e+00	2.771	0.005592	**
F30	3.697e+00	5.825e-01	6.347	2.19e-10	***
F33	2.423e-04	1.436e-04	1.687	0.091636	.
F34	-9.036e-06	4.331e-06	-2.086	0.036945	*

挥了重要作用。

3.3 垃圾评论识别效果及其分析

本文通过建立分类模型检测第一类垃圾评论，由于人工标记训练集是比较困难的，而在 2.1.1 节中提到的三种类型重复评论几乎可以确定为垃圾评论。因此，本文将所有重复评论归为正类，其他剩下的评论归为负类，以此来建立模型。同时，使用十倍交叉验证来获得实验结果，针对不同特征集合的实验结果如表 3 所示。

表 3 使用不同特征构建的逻辑回归模型 AUC 值

特征的使用	AUC/%
所有特征	83.8
所有评论特征	76.2
所有评论者特征	77.3
除去 feedbacks 之外的特征	83.7
所有文本特征	70.1
'***'特征	82.2
'***'特征和'**'特征	84.7
'.'以上的所有显著性特征	84.6

从实验结果可以看出：

(1)使用所有特征 AUC 为 83.8%，考虑到负类样本中的许多非重复评论也有可能是垃圾评论，该 AUC 值已经相当高了。

(2)除去 feedbacks 之外的特征 AUC 值为 83.7%，证明 feedbacks 在垃圾评论检测中作用不明显。

(3)单独文本特征的 AUC 值只有 70.1%，说明不能单独使用文本特征进行垃圾评论的识别。

(4)'***' 特征 AUC 值为 82.2%，比所有特征只差 1.6%，而'***'特征和'**'特征 AUC 值最高，甚至高于所有特征的 AUC 值达到 84.7%。

综上所述，本文提出的以显著性特征构建的模型更加稳定，不仅减少了计算量，而且能够达到和所有特征同样的效果。

当然，利用重复和非重复评论建立逻辑回归模型不只是为了检测重复评论，因为重复垃圾评论的识别可以通过简单的内容比较检测到(见 2.1.1 节)，本文真正的目的是用该模型来识别第一类型垃圾评论中的非重复评论。上述实验结果证明了模型可预测重复评论，为了进一步确认它的可预测性，需要证明它也可以预测那些非重复的垃圾评论。

为此，本文将通过人工检测查看许多排名很高的非重复评论是否是真正的垃圾评论。首先，对负类测试样本(非重复)按照概率进行排列；然后，对排名较高的评论进行人工标注，看它们是否为垃圾评论。人工标注采用投票的方式来完成。实验结果如表 4 所示，其中第二行为应用所有特征的负类样本排列后检测的结果，第三行为应用'***'特征和'**'特征的结果。可以看出使用所有特征以及'***'特征和'**'特征能够识别的垃圾评论数量都较少，这是由于许多有经验的垃圾评论者能够很好地掩饰他们的行为，使判别变得异常困难。但实验结果表明，本文提出的方法是有效的，可以应用更少的显著性特征来识别产品垃圾评论。

表 4 负类样本中排名靠前的产品垃圾评论

排名	1~15	16~30	31~45	46~60	61~75	75~90
所有特征	11	12	8	10	6	9
显著性特征	13	9	9	8	7	8

以过多的产品评论特征建立的逻辑回归模型存在模型不稳定和计算量大的问题。对特征进行显著性检验能有效解决该问题。本文对 JINDAL N 等人提出的方法进行研究，并分析了存在的问题，提出利用更为合理的显著性特征建立模型，从而提高了模型质量。新的模型更加稳定，使得计算量大大减少。实验通过亚马逊数据集，验证了本文方法的有效性。未来的工作将致力于通过混合各种算法和分类器来提高算法精度。

参考文献

[1] JINDAL N, Liu Bing. Review spam detection[C]. Proceedings of the 16th International Conference on World Wide Web, 2007: 1189-1190.

[2] JINDAL N, Liu Bing. Opinion spam and analysis[C]. Proceedings of the International Conference on Web Search and Web data mining, 2008: 219-230.

[3] HU M, Liu Bing. Mining and summarizing customer reviews [C]. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004: 168-177.

[4] PANG Bo, LILLIAN L E. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval. 2008, 2(1-2): 1-135.

[5] BRODER A Z.On the resemblance and containment of documents[C].Proceedings of the Compression and Complexity of Sequences.1997:21-29.

(收稿日期:2012-07-17)

作者简介:

吴敏,男,1988年生,硕士研究生,主要研究方向:文本倾向性检索与研究、信息安全。

何珑,男,1971年生,高级工程师,硕士生导师,主要研究方向:数据库、网络系统、信息安全。