IJCSE 国际计算机科学与工程开放获取期刊

研究论文 第 7 卷,第 5 期,2019 年 5 月 电子期刊号:2347-2693

使用新闻文章和博客进行句子级情感分析机器学习技术

维沙尔·希尔萨特1* ， Rajkumar Jagdale2 , 坎查申德3 , 萨钦·N·德什穆克4 , 苏尼尔·卡瓦莱5

1部印度巴巴萨海布·安贝德卡尔·马拉特瓦达大学计算机科学与信息技术系,奥兰加巴德 431004
2部巴巴萨海布·安贝德卡尔·马拉特瓦达大学统计学系,奥兰加巴德 431004,印度

*通讯作者: vss.csit@gmail.com

DOI:https://doi.org/10.26438/ijcse/v7i5.16 | 在线获取:www.ijcseonline.org

接受日期:2019年5月7日,发布日期:2019年5月31日

摘要 如今,情绪分析在文本挖掘中起着非常重要的作用。本质上,网络挖掘是数据挖掘领域中一个非常广泛的领域,用于提取文本的情绪。识别文本数据的情绪是一项非常具有挑战性的任务。本研究侧重于从新闻文章和博客中进行句子级否定识别和计算。分析通常使用两步方法,即预处理和后处理。预处理包括删除停用词、删除标点符号、删除数字、删除空格等任务。后处理包括从文本中识别情绪和计算分数。本研究分析了支持向量机、朴素贝叶斯对在线收集的数据集的性能。

关键词:情感分析,支持向量机,朴素贝叶斯,机器学习算法

一、引言

情感分析是自然语言处理和文本挖掘的应用之一,用于从文本中提取主观信息,也可用于提取文本数据的上下文极性、情感。数据提取可以从各种数据源（如推文、博客、社交媒体和在线新闻文章)中进行挖掘。

情感分析分为三个层次:文档级别,句子级别,实体和方面级别。

文档级别方法分析整个文档的极性。文档通常包含一个项目的评论;因此系统将计算或表达关于项目的整体极性。句子级别处理和分析每个语句以确定极性并给出每个句子的极性。第三级即实体和方面级别,这是情绪分析中最重要的级别。与之前的级别相比,这一级别是基于特征的情绪分析,有助于找出实体及其方面的情绪。本文借助机器学习算法提出了否定在新闻文章和博客中的作用。如今,新闻文章和博客是最重要的平台,允许用户表达对几个问题的个人看法,这可能是

涉及政治、社会责任以及国家或国际问题等。网络上有大量文本形式的数据,情感分析的目的是找出用户意见的极性。情感分析

我们可以轻松预测个人或群体对上述特定问题的看法。

二.相关工作

情感分析领域的研究人员开发并应用了各种算法来预测新闻文章和博客中文本的情感。这些可能基于自然语言处理 (NLP)、基于模式的技术和机器学习算法,例如朴素湾 (NB)、支持向量机和随机森林。一些研究人员使用了无监督和半监督学习技术。

情感分析领域的研究人员已经开发并应用了各种算法来预测新闻文章和博客文本的情感。这些算法可能基于自然语言处理 (NLP)、基于模式的技术和机器学习算法,例如朴素贝叶斯 (NB)、支持向量机和随机森林。一些研究人员使用了无监督和半监督学习技术。

M. Thelwall 和 K. Buckley [1] 表明,在情感分析中,存在三种方法,例如基于机器学习的方法、基于词典的方法和语言分析。 Jagdale, R. S [2] 区分了 Twitter 上的不同事件分析并计算情绪

每个事件的极性。L. Tan,J. Na,Y. Theng [3] 的语言学方法是分析文本方向,语言学方法使用单词或短语的句法特征、否定和文本结构。自然语言文本的情感分析是一个广泛且不断扩展的领域。文本可能包含主观和客观情感。Wiebe [4] 将主观文本定义为某人的观点、情感、情绪、评价、信仰和猜测的语言表达。在她的定义中,作者受到了语言学家 Ann Ban field [5] 的作品的启发,她将主观定义为一个从人物的角度出发并呈现私人状态（不对客观观察或验证开放)的句子,由 Quirk [6] 定义,是一种体验,持有一种态度,可以选择对某个对象。

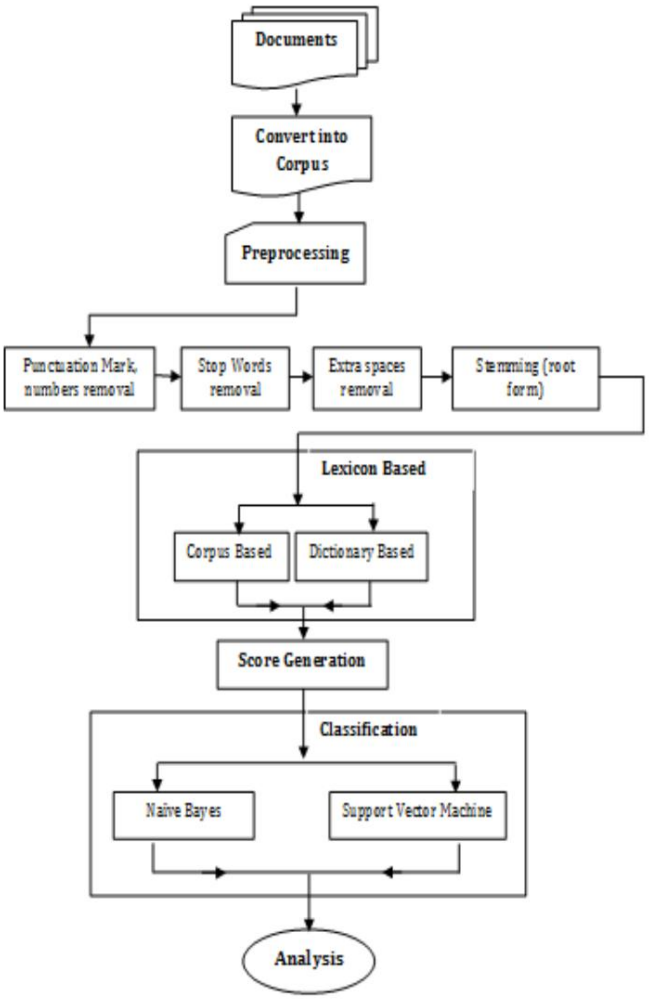
刘兵[7]将客观文本定义为关于实体、事件及其属性所表达的事实。Esuli 和 Sebastiani[8]将情感分析定义为信息检索和计算语言学交叉领域的一门新兴学科,它关注的不是文档的主题,而是文档所表达的观点。

Shoukry [9] 展示了一款用于阿拉伯语情绪分析的应用程序,并对阿拉伯语推文进行了情绪分类。研究人员对收集到的推文进行了分析,以确定其极性。他们的研究提出了一种混合系统,该系统使用了 ML 方法中识别出的所有特征,以及 SO 方法中的情绪词典,准确率和召回率为 80.9%,而精确度和 F 度量为 80.6%。Alexandra Balahur [10] 的这篇论文研究了在三个层面上定义任务的重要性的重要性。他们对目标文本进行了分析,并区分了好消息和坏消息内容。

三 .方法

预处理在情感分析中起着非常重要的作用,预处理的作用是从文本中删除不需要的数据。这种类型的数据不包含任何重要信息。在这项工作中,我们使用 BBC 新闻文章数据集和多感官数据集进行实验。

每篇在线文本都包含 HTML 标签、脚本和广告等信息。此步骤有助于清理和准备数据以供后期处理。本研究使用多传感器新闻文章和 BBC 新闻文章数据集 [12, 13]。多传感器新闻文章数据集包含 12,073 篇文档,类别包括经济、健康、生活方式、自然环境、政治和科学技术,BBC 数据集包含 2225 篇文档,类别包括商业、政治、娱乐、体育、技术等。



数字。 1 提议的方法

机器学习算法

这项工作使用机器学习算法进行分类。使用的分类器是朴素贝叶斯 (NB) 和支持向量机 (SVM)。朴素贝叶斯分类器根据已经发生的另一个事件的概率来找到发生的概率。NB 分类器对于线性可分问题甚至非线性可分问题都表现得非常好,并且表现相当好 [14]。

$$(|) \qquad \frac{(|)\ ()}{( )} \qquad ()$$

上式(1)中,A为文本的情感,B为文本,P(A|B)为类别的后验概率,P(A)为类别的概率。 P (B|A) 是似然,P (B) 是预测变量的先验概率。SVM 是非概率算法,用于分离数据

顺序和非顺序 [15]。它主要用于文本分类,在高维特征空间中表现良好。SVM 表示实例点空间,映射使得不同类的实例尽可能被清晰的边缘分开 [16]。

五、结果与讨论

本节介绍后处理工作。如上所述,使用了两个数据集,即 BBC 和 Multisensor。提议的方法采用五个步骤。第一步执行数据清理,从数据中删除 URL、停用词、标点符号、删除空格和数字。此外,删除数字是一个重要的步骤,因为数字很少代表情绪,因此并不重要。下一步是将文档转换为小写以保持统一。词干提取用于将单词的词根形式更改为单个单词,例如“Connection”、“Connecting”、“Connected”,即“Connect”。术语文档频率描述了文档中出现的术语的频率,其中输出中的行被假定为集合,列被假定为相关术语。情感识别是一项主要任务,为了实现这一目标,我们使用了sentimentr和syuzhet包以及基于字典的方法和基于词典的方法。Sentimentr 包使用 10 个词典来进行具有 11 个参数的情感识别,并计算句子的极性,并利用情感词典来标记极化词。该包使用以下公式计算情绪。

Σ(( ) ( ))
( ) Σ(( ) ( ))
)
(Σ(( ))

Σ(| |
| | *-1)

(Σ)

√ ( )

Sentimentr 包使用上述公式来计算文本的情感[6]。 Syuzhet 包使用三种方法来计算情感:NRC、Bing 和 Afinn,这三种方法给出了不同的结果[17]。然而,在实验中,我们发现与其他包相比,Sentimentr 包最适合句子级别,因为 Sentimentr 包借助更多参数来计算每个句子的极性,例如正面词、负面词、Downtowners、放大器、反放大器、逆向连词ETC。

表 1. 显示了 Sentimentr 和 Syuzhet 包及其参数和计算方法的比较

的情绪。而表 2 显示了 BBC 数据集的类别极性。在给定的表中,有商业、娱乐、政治、体育和技术五个类别。为了计算句子的极性,我们使用了sentimentr包。在表 3 中,我们使用了两种机器学习技术:朴素贝叶斯和支持向量机,计算了准确度、精度和 f 分数。这里,娱乐类别通过使用朴素贝叶斯获得了更高的准确性,支持向量机为商业类别提供了更高的准确性。这两种机器学习算法都是情感分析中最流行的概率和非概率算法。之后表 4 显示了 BBC 数据集的混淆矩阵,该矩阵对结果进行对角分类。在混淆矩阵中,结果应该是对角线的。在第二个实验中,我们使用多传感器博客数据集,而表 5 显示了积极、消极和中性的结果。在这个实验中,我们使用sentimentr包来计算博客的极性。对于统计测量,还计算平均误差、均方根误差、平均绝对误差和平均绝对平方误差。最后我们知道,我们得到的这些结果是通过机器学习算法和统计测量证明的。

桌子。 1 Sentimentr 和 Syuzhet 包的比较

序号	的名字 包裹	参数和方法
1	情感	正面词,负面词,市中心人, 放大器、去放大器、对抗 连词
2	修泽特	NRC、Bing 和 Afinn

表 2. BBC 数据集的类别极性

先生。 不	的名字 类别	文章			
		全部的	正面 负面 中性		
1	商业	510	262	214	三十四
2	娱乐	401	136	244	21
3	政治	第417章	210	190	17 号
4	运动	511	151	327	33
5	科技	401	136	244	21
	全部的	2,240 人	895	1219	126

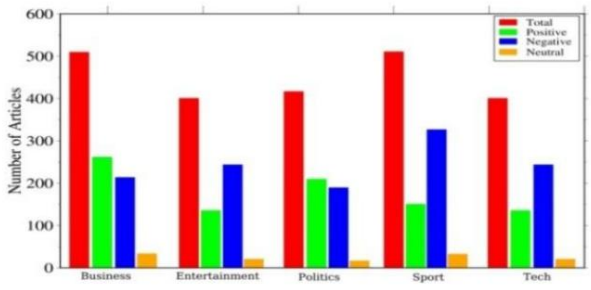


图 2:BBC 数据集的类别极性

表 3 Naïve Bays 的比较分析

数据集	朴素贝叶斯		
	准确性	精确	F-分数
商业	92.63	89.76	91.32
娱乐	96.46	94.80	97.33
政治	93.33	88.88	93.33
运动	93.00	90.74	95.14
技术	96.46	94.80	97.33

表4 SVM比较分析

数据集	支持向量机		
	准确性	精确	F 分数
商业	82.60	79.67	89.34
娱乐	69.91	68.22	84.39
政治	94.16	89.06	94.21
运动	69.23	69.01	81.66
技术	69.91	68.22	81.11

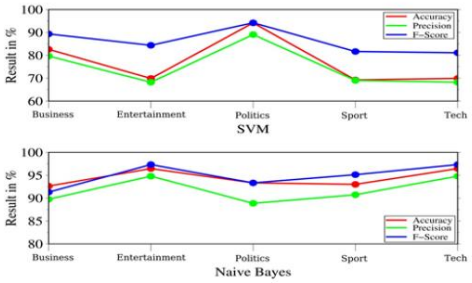


图 3:Naïve Bays 与 SVM 的比较分析

表 5.混淆矩阵

A	b	C	d		分类为
1	0	0	0	A	娱乐
0	1	0	0	乙	政治
0	0	1	0	C	运动
0	0	0	1	d	技术

表 6. 多传感器数据集的类别极性

长者 不	的名字 类别	博客			
		全部的	积极的	消极的	中性的
1	经济 商业 金融	3689	2488	第371章	第830章
2 健康		326	198	47	81
3	生活方式 闲暇	3353	2471	第387章	第495章
4	自然 环境	990	第673章	162	155
5	政治	第561章	315	87	159
全部的			2386	2047	97

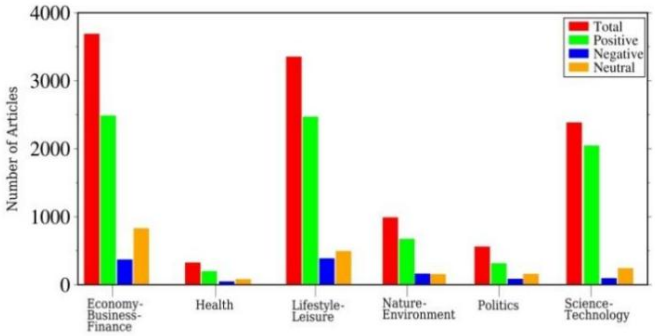


图 4:MULTISENSOR 的类别极性

数据集

表 7:多传感器训练统计测量

数据集

数据集	训练			
	梅			硕士
经济_商业_金融	-4.752	0.122	0.090	1.000
健康	1.369	0.273	0.195	1.000
生活方式_休闲	-1.781	0.178	0.128	1.000
自然_环境	4.158	0.168	0.128	1.000
政治	2.232	0.095	0.074	1.000
科学技术	-1.110	0.157	0.110	1.000

表 7:多传感器测试统计测量

数据集

数据集	测试			
	我 RMSE	梅 MASE		
经济_商业_金融	-8.429	0.115	0.076	0.837
健康	-4.581	0.121	0.098	0.503
生活方式_休闲	-4.891	0.203	0.142	1.111
自然_环境	3.219	0.205	0.142	1.106
政治	1.757	0.102	0.075	1.014
科学技术	-4.571	0.130	0.097	0.883

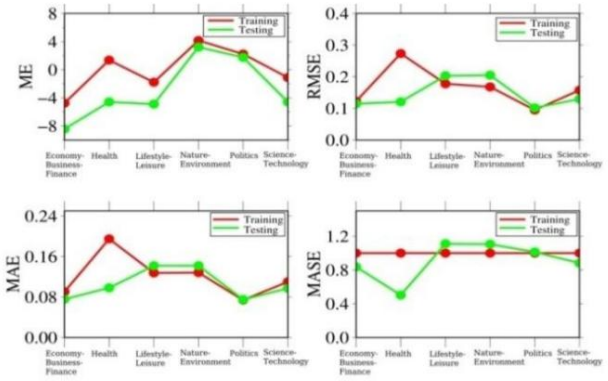


图 5:多传感器数据集统计测量的比较分析

VI. 结论和未来工作

在这项工作中,我们对新闻文章和博客进行了句子级情感分析。进行了实验工作以计算新闻文章和博客的极性。结果显示了类别句子级极性。分类建议工作使用了朴素贝叶斯、支持向量机和随机森林。其中朴素贝叶斯的准确率达到 96.46%,而支持向量机算法的准确率达到 94.16%。通过这些结果,我们知道朴素贝叶斯与支持向量机相比取得了更好的结果,因为与支持向量机相比,朴素贝叶斯研究了给定句子中每个单词及其特征的概率。

未来的工作将集中于来自新闻文章和博客网站的其他数据的其他分类机器学习技术。

致谢

我非常感谢奥兰加巴德巴巴海布·安贝德卡尔马拉特瓦达大学计算机科学与信息技术系和德里大学拨款委员会以拉吉夫·甘地国家奖学金 (SRF) 的形式为我提供与我的研究工作相关的所有便利。

参考

[1] M. Thelwall、K. Buckley、G. Paltoglou，“推特事件中的情绪”，美国信息科学与技术学会杂志 62(2), (2011) 406 418。

[2] Jagdale, RS、Shirsat, VS、Deshmukh, S. N，“使用开源工具对 Twitter 事件进行情绪分析”。《国际计算机科学与移动计算杂志》（IJCSMC），第 5 卷,第 4 期,2016 年 4 月,第 475-485 页。

[3] L. Tan, J. Na, Y. Theng, K. Chang，“使用语言学方法进行句子级情感极性分类,数字图书馆:为了文化遗产”，知识传播与未来创造（2011)77-87。

[4] Wiebe,J.，“追踪叙述中的观点。”计算语言学,20,1994 年。

[5] Banfield,A.,《无法言说的句子:小说语言中的叙述和表述》,Routledge and Kegan Paul,1982 年。

[6] Quirk,R.，“英语综合语法”，
朗文出版社,1985年。
[7]刘兵，《情感分析与主观性》,自然语言处理手册,第二版,2010年。

[8] Esuli, A. 和 F. Sebastiani，“SentiWordNet:一种可公开获取的观点挖掘资源”，《第六届语言资源与评估国际会议论文集》,LREC 2006,意大利,2006 年。

[9] Shoukry, Amira,协作技术和系统（CTS）,2012 年国际会议技术和系统,2012 年 5 月 21-25 日,第 546 – 550 页

[10] Alexandra Balahur,Ralf Steinberger，“重新思考新闻中的情绪分析”,理论到实践和返回”,欧盟委员会,联合研究中心,阿利坎特大学软件和计算系统系。WOMSA,第 1 页-
12,2009。

[11] Ye, Q., Zhang, Z., Law, R.:通过监督机器学习方法对旅游目的地在线评论进行情感分类。专家系统。应用。 36, 6527–6535 (2009)

[12] D. Liparas、Y. Hacohen-Kerner、A. Moumtzidou、S. Vrochidis 和 I. Kompatsiaris，“使用随机森林和加权多模态特征进行新闻文章分类”,第三届开放跨学科 MUMIA 会议和第七届信息检索设施会议（IRFC2014）,丹麦哥本哈根,2014 年 11 月 10 日至 12 日。

[13] D.格林和P.坎宁安。 “内核文档聚类中对角优势问题的实用解决方案”,Proc。
ICML 2006。

[14]叶倩文、张哲瀚、罗瑞:通过监督机器学习方法对旅游目的地在线评论进行情感分类。专家系统应用,第 36 卷,第 6527–6535 页 (2009)

[15] Bhumika, M.、Jadav, V.、Vaghela, B.:基于特征选择和语义分析的支持向量机情感分析。Int. J. Comput. Appl. 146(13)（2016 年）。

[16] BholaneSavita,D., Deipali, G.:使用支持向量机对 Twitter 数据进行情感分析。国际。 J. 计算机。科学。趋势技术。 4(3) (2016)

[17]林克,台湾（2018）。情感:计算文本极性情感版本2.6.1。 http://github.com/trinker/sentimentr

[18] Jockers ML (2015)。Syuzhet:从文本中提取情感和情节弧。https://github.com/mjockers/syuzhet。

作者简介

Sachin N. Deshmukh目前担任奥兰加巴德马拉斯瓦达大学 Babasaheb Ambedkar 博士计算机科学与 IT 系教授,在研究生（M. Tech、M.Sc 和 MCA)和研究生教学方面拥有约 24 年的经验BE、B. 技术课程。在国内外知名期刊和会议上发表研究论文80余篇。大学当局还担任孟买印度理工学院口语辅导项目主任（大学网络信息中心)主任、职业教育培训中心主任、首席协调员。他还参与了教资会和AICTE的研究项目。除了大学之外,还曾在新德里 AICTE 担任副主任（电子政务），并在 COEP Pune on Lien 担任副教授。他是 EQASA 成员、Manber NAAC 同行团队成员、UGC 委员会成员、AICTE-SCSC 和 AICTE-SCAC 成员、西班牙圣地亚哥 PEIN 研究员、IETE 研究员。他的研究领域是文本挖掘、社交媒体数据分析、情感分析与观点挖掘、内涵挖掘。

Sunil Kawale目前担任奥兰加巴德巴巴萨海布·安贝德卡尔马拉特瓦达大学统计学系教授,拥有约二十年研究生（统计学硕士)教学经验。他的研究领域是运筹学、随机过程、计算机编程和应用、数据挖掘。

维沙尔·S·希尔萨特 (Vishal S. Shirsat)荣获菲尔硕士学位。（计算机科学)毕业于奥兰加巴德巴巴萨海布·安贝德卡尔·马拉特瓦达大学计算机科学与信息技术系,目前在同一系攻读博士学位,并因其研究工作获得国家奖学金。他的研究领域是情感分析和观点挖掘。

Rajkumar S. Jagdale已获得理学硕士学位。（计算机科学)毕业于奥兰加巴德巴巴萨海布·安贝德卡尔·马拉特瓦达大学计算机科学与 IT 系,目前在同一系攻读博士学位,并因其研究工作获得 DST Inspire 奖学金。他的研究领域是情感分析意见挖掘。

Kanchan Shende已获得哲学硕士学位。（计算机科学)毕业于奥兰加巴德巴巴萨海布·安贝德卡尔·马拉特瓦达大学计算机科学与 IT 系,目前在同一系攻读博士学位,并因其研究工作获得国家奖学金。她的研究领域是海洋学、计量学、遥感和 GIS。