

微博立场检测研究综述

周耘立,李珊珊

(四川大学计算机学院,成都 610065)

摘要:

微博立场检测是一个新兴的研究课题,具有很大的研究价值和应用价值。鉴于此,该研究课题受到越来越多国内外研究机构的重视。首先针对立场检测主流方法和前沿进展进行介绍;接着介绍国内外主流的评测会议;最后,总结并展望立场检测技术的发展趋势。

关键词:

情感倾向性分析;立场检测;微博

0 引言

立场检测任务就是自动的检测一段文本的作者对于给定的目标对象他的立场是支持、反对还是中立。近几年,已经在立场检测建模方面进行了积极的研究工作^[4-5,10]。Faulkner^[4]使用目标信息和维基百科链接关系来自动的识别学生论文中的目标立场。由于微博文本简短、非正式、充满错误拼写、简写以及俚语等特点,给研究界带来了新的机遇和挑战^[2,5]。微博立场检测任务旨在从一段微博中检测作者对于给定目标的立场,而不依赖于在线讨论中经常出现的会话结构。

立场检测与情感分析研究相关,但是两者有着显著的差别。立场检测通常可以看做是观点挖掘和情感分析的一个子任务。在情感分析中,我们通过对文本中的观点、态度的分析来判断文本所表达的情感是正面的还是负面的。然而,在立场探测中,系统需要确定作者对于给定目标的偏好,是支持还是反对,或者是中立。这个目标可能并没有在文本中显式的出现,而是被暗指。同时这个文本可能对其他实体对象表达观点或者情感^[6]。例如,考虑表1中的目标和微博文本。

微博中并没有直接提及目标对象,也没有直接表对目标表达情感倾向,但是我们可以推断出微博作者很可能是不支持目标对象的。立场检测系统需要能自

动识别出不存在于焦点文本中的相关信息。自动的立场检测技术有着多方面的应用,如信息检索、文本摘要和文本蕴涵等^[11]。

表1 样本数据

目标	全面放开二孩政策
微博	我们都是可以随便安排的机器么

1 技术方法

针对微博进行情感倾向性分析研究技术主要有三大类:第一类是采用情感词典与规则结合的方法^[1,3],根据文本中所包含的情感词的个数与情感词的倾向来进行立场检测;第二类是采用机器学习的方法,选择文本中的一些特征,标准训练集和测试集,使用朴素贝叶斯、SVM^[7,11]等分类器进行立场检测;第三类是规则与机器学习结合的方法。

微博立场检测主要是考虑对于给定的目标对象微博作者的立场倾向,这个对象可能是事件、人或者物体。目前主要的研究方法是基于机器学习的方法。按照机器学习类型划分,我们可以将机器学习方法分为两类:第一类是有监督学习方法;第二类是无监督学习方法。

1.1 有监督方法

有监督方法将立场检测问题看做是一个分类问

题,使用机器学习方法从标注数据中学习分类器。当前大多数的语言学研究者把有监督的立场检测问题看做是文档水平的立场分类问题,使用各种词汇、语法和语义特征来识别特定目标的立场。针对微博文本的特点,Patra^[12]使用了 SentiWordNet、NRC Emotion Lexicon、Hashtag Emotion Lexicon 三个情感词典来提取词级别的特征,同时借助于 RitaWordNet 外部词典资源来创建目标相关主题词袋。此外,还结合依存分析和手工规则构建特征,最后使用支持向量机来训练分类器。Krejzl^[13]通过使用 hashtags 的情感极性、n 元语法、微博长度以及强烈表明一个特定立场的立场领域词典等构造特征,训练一个最大熵分类器。Elfardy^[14]试图探究词法和语法特征能多大程度上来帮助识别微博立场,提出在有监督框架下,使用词汇、语义字典和潜在语义情感来识别情感立场。

1.2 无监督方法

有监督方法主要缺点是依赖于人手工标注数据,这是一项十分费时费力的工作。当没有办法获取大规模的标注语料时,人们提出了无监督学习方法,仅使用非标注数据进行建模。Turney 通过计算文中候选单词与种子情感词之间的点互信息来计算文本的情感倾向性^[7]。由于在线辩论语言使用上的复杂性,辩论双方常常使用先褒后贬的方式来表达自己的观点,Somasundaran^[9]提出通过构建约束条件利用整数线性规划进行建模,帮助识别作者的立场。立场检测中目标对象很可能并没有出现在微博中,Augenstein^[13]等人通过从无标注的微博数据中使用词袋自编码模型学习特征,提高模型在目标对象未出现微博上的泛化能力从而提高了立场检测的效果。

2 评测数据和评测指标

微博用户产生的大量评论性数据潜藏着巨大的研究价值和商业价值^[7]。微博立场检测首先引起国际语义评估研讨会 (SemEval-20161) 的关注,2016 年发布了“Detecting Stance in Tweets”评测任务。国内第五届自然语言处理与中文计算会议(NLPCC20162),发布了“中

文微博立场检测”评测任务。国内外研究机构都关注在两个任务上,第一个任务是有监督方法的立场探测;第二个是无监督方法的立场检测。下面针对 SemEval-2016 评测数据进行介绍。

SemEval-2016 立场检测任务中,评测数据包含五个目标对象分别是“Atheism”,“Climate Change is a Real Concern”,“Feminist Movement”,“Hillary Clinton”,“Legalization of Abortion”,所有的数据都标注了立场。表 2 所示是训练集上样本的统计。

表 2 训练集上样本数据统计

	训练集				
	Abortion	Atheism	Climate	Clinton	Feminism
Favor	105	92	212	112	210
Against	334	304	15	361	328
None	164	117	168	166	126
All	603	513	168	166	126

精确率、召回率和 F 度量值三种评价指标在自然语言处理领域被广泛使用。其中精确率和召回率从不同的角度度量了分类器的好坏,F 度量值综合了精确率和召回率两种指标^[3]。

表 3 分类器输出结果

	识别为正例	识别为负例
实际为正例	TP	FN
实际为负例	FP	TN

精确率(P)、召回率(R)、F 值定义如下:

$$P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN} \quad F = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

3 结语

微博作为主流的社交媒体具有重大的研究价值和商业价值,微博立场探测作为新兴的研究课题,得到国内外研究学者的广泛重视,并成为研究热点。本文主要从立场检测前沿进展、主流方法以及相关评测进行介绍。目前基于神经网络的方法也开始在立场检测任务中展露光芒,相信在未来会有更多的进展。由于微博文本本身的一些随意性、不规范表达增加了立场检测的

参考文献:

- [1]赵妍妍,秦兵,刘挺.文本情感分析[J].软件学报,2010,21(8):1834-1848.
- [2]张剑峰,夏云庆,姚建民,等.微博文本处理研究综述[J].中文信息学报,2012,26(4):21-27.

- [3]奉国和. 文本分类性能评价研究[J]. 情报杂志, 2011(8):66-70.
- [4]Faulkner A. Automated Classification of Stance in Student Essays: An Approach Using Stance Target Information and the Wikipedia Link-Based Measure[C]// 2014.
- [5]Kiritchenko S, Zhu X, Mohammad S M. Sentiment Analysis of Short Informal Text[J]. Journal of Artificial Intelligence Research, 2014, 50:723-762..
- [6]Mohammad S M, Sobhani P, Kiritchenko S. Stance and Sentiment in Tweets[J], 2016.
- [7]杨经, 林世平. 基于 SVM 的文本词句情感分析[J]. 计算机应用与软件, 2011, 28(9):225-228.
- [8]陈渊, 林磊, 孙承杰, 等. 一种面向微博用户的标签推荐方法[J]. 智能计算机与应用, 2011, 01(5):21-26.
- [9]Somasundaran S, Wiebe J. Recognizing Stances in Online Debates[J], 2009.
- [10]Mohammad S, Kiritchenko S, Sobhani P, et al. SemEval-2016 Task 6: Detecting Stance in Tweets[C]. International Workshop on Semantic Evaluation, 2016.
- [11]Patra B G, Das D, Bandyopadhyay S. JU_NLP at SemEval-2016 Task 6: Detecting Stance in Tweets using Support Vector Machines [C]. Semeval, 2016.
- [12]Krejzl P, Steinberger J. UWB at SemEval-2016 Task 6: Stance Detection[C]. International Workshop on Semantic Evaluation, 2016.
- [13]Augenstein I, Vlachos A, Bontcheva K. USFD at SemEval-2016 Task 6: Any-Target Stance Detection on Twitter with Autoencoders [C]. International Workshop on Semantic Evaluation, 2016.
- [14]Elfardy H, Diab M. CU-GWU Perspective at SemEval-2016 Task 6: Ideological Stance Detection in Informal Text[C]. International Workshop on Semantic Evaluation, 2016.

作者简介:

周耘立(1990-),男,四川成都人,硕士研究生,研究方向为数据挖掘

李珊珊(1989-),女,江苏徐州人,硕士研究生,研究方向为自然语言处理

收稿日期:2016-11-15

修稿日期:2017-01-05

Summary of Micro-Blog Stance Detection

ZHOU Yun-li, LI Shan-shan

(College of Computer Science, Sichuan University, Chengdu 610065)

Abstract:

Micro-blog stance detection is gaining importance due to various academic and commercial perspectives. For that reason, the subject of research receives increasingly attention at home and abroad in research institutions. Summarizes the author first the main stream stance detection approaches and the most recently research progress. Then introduces the main stream of evaluation meetings at home and abroad, finally, summarizes and forecasts the development trend of stance detection technology.

Keywords:

Sentiment Analysis; Stance Detection; Micro-blog