廣東工業大學

本科毕业设计（论文）

基于 BERT 模型的文本情感分析

学　　院	计算机学院
专　　业	网络工程
	（高速网络与云计算方向）
年级班别	2016 级（2）班
学　　号	3116004982
学生姓名	赵舒宇
指导教师	王卓薇

2020 年 2 月

摘　　要

TODO

关键词: BERT，情感分析

Abstract

TODO

KEY WORDS: BERT, Sentiment Analysis

目　　录

1 引言

1.1 研究背景及意义

1.1.1 研究背景

得益于摩尔定律[1]，生产电子元件的成本越来越低，电子产品的性能越来越好；同时，网速的高速发展也促进了移动互联网的进步。这两者都使得我国网民数量持续增长，用户在互联网上扮演的角色已经从单纯的信息接收者，变成现在信息的生产者。在社交媒体上，电商网站上，数亿的用户产出海量的数据，且这些数据是以指数级增长的，人工分析利用这些数据需要耗费大量的时间和人力，属于不可能的任务。但这些数据有着重要的社会价值和商业价值，如在诸如微博的这类社交媒体上，分析用户针对不同社会话题发布的微博内容，可以有效的监控网络舆情[2]；而在淘宝这类电商平台上，分析用户给予不同商品的评价，商业公司可以快速了解这一商品的受欢迎程度。所以，如何通过自动化工具正确，快速分析利用这些数据，成为当今计算机科学研究领域重要的话题。

文本情感分析（Text Sentiment Analysis）作为自然语言处理（Natural Language Processing）领域的基本研究分支之一，主要任务是对文本的主客观性，观点，情绪，喜好的检测，分析，挖掘。文本情感分析作为多学科交叉研究领域，涉及语言学，统计学，机器学习，数据挖掘，人工智能等多个领域。近年来，随着机器学习的发展，学术界在情感分析技术上取得长足进步，工业界也开发许多情感分析技术落地的项目。自然语言处理于二十世纪中叶于美国兴起，主要应用场景都是针对英语。中文自然语言处理起步晚于英语，同时，中文自然语言处理也与英语有许多不同，主要表现在以下几个方面：①中文需要进行分词；②中文没有明显的屈折变化（时态，单复数等）；③中英文句法结构上存在不同。但随着技术的发展和数据的爆炸式增长，作为中文自然语言处

理的一个子类，提取分析非结构化文本的情感分析技术也愈发成熟。

1.1.2 研究意义

在像微博这样的社交媒体上，大量用户针对不同话题广泛发表自己的见解。这些文本通常数量庞大，依靠人力根本无法进行分析。但这些数据又有着重要的意义[2-3]。利用这些数据，可以迅速准确的把握微博平台上重要事件的情感倾向，有效的进行舆情监控，对政府，商业公司维持舆论稳定有着极大帮助，这时就彰显出情感分析技术的重要性。

而对于淘宝，京东这样的电商网站，它们拥有着大量的商家和消费者用户。消费者们乐于对自己喜欢的商品给予好评，对于不满意的产品也会留下差评。而这些评论对商家的发展十分重要[4]。根据评论，商家可以了解到顾客对于产品的喜好程度和产品的不足之处。针对这些客户意见，商家可以有针对性的对商品进行改善，提高客户满意度的同时，吸引更多的潜在客户，提高商业利润，而情感分析技术正能有效的帮助商家实现这些目的。

鉴于上述情况，本文选取微博文本数据和电商评论数据，针对这两个数据集进行情感分析方法研究。尝试不同的方法来研究情感分析技术，并设计实验比较各种方法的分类结果，发掘出性能较好的分类模型，进而改善情感分类效果。

1.2 国内外研究现状

文本情感分析的核心问题是情感分类，主要研究任务分类两类，一是区分主客观文本，降低因客观信息对情感分析性能的干扰；二是对主观性文本进行情感分类[5]。根据对文本划分的粒度不同，又可分为对词，句，篇章的情感分析。根据情感划分的粒度不同，可分为：①二元分类，包括消极态度和积极态度。②多元分类，根据人类的多种情感进行进一步细分，包括“快乐，悲伤，褒扬，贬斥，信息，意外”等十大类[6]。

而对于主观文本情感分类，现在流行的方法包括以下三种：

1) 基于情感字典的方法。这一方法最符合人类的直觉，首先构建包含大量基本词汇的字典，字典中的词已经标记好词的类别。如“喜欢”会被标记为积极，而“讨厌”会被标记为消极。之后将输入的文本与字典进行匹配来判断其情感极性。这一方法有着局限性，通常与其他方法一同使用。

2) 基于机器学习的方法。这一方法关键点在于选择有效特征组合利用分类器进行情感分类[7]。这一方法可以取得不错的成果，但对数据集要求很高，往往需要大量人工标注，成本不菲。

3) 基于深度学习的方法。这一方法通过建立深度神经网络，模拟人脑结构，在处理情感分类这种问题上有着优异的表现，是目前情感分析领域的主流方法。

1.2.1 国外研究现状

对于文本主客观分类，Wiebe 等[8] 人于 1999 年就将代词，形容词，基数词等作为特征值，设计了对文本主客观进行分类的朴素贝叶斯分类器 (Naive Bayes Classfier)。在 2003 年，Rlloff 与 Wiebe[9] 提出 bootstrapping 文本主客观分类算法。该算法提高分类召回率的同时，只损失部分的精度。之后，Wiebe 等[10] 人还针对篇章粒度下的文本主客观分类进行了研究。在情感字典技术方面，Tumey[11] 设计了点互信息的方法（Point Mutual Information, PMI) 来计算两个词之间的语义相关性。通过计算目标词与情感词之间的 PMI，用语义倾向 (Semantic Orientation, SO) 来表示该词的情感极性。公式如下：

$$\mathrm{PMI}\left(word_1, word_2\right) = \log_2\left[\frac{\mathrm{p}\left(word_1 \& word_2\right)}{\mathrm{p}\left(word_1\right)\mathrm{p}\left(word_2\right)}\right] \tag{1.1}$$

$$\begin{aligned}\mathrm{SO}\left(\mathrm{phrase}\right) = {} & \mathrm{PMI}\left(phrase, \text{“excellent”}\right) \\ & - \mathrm{PMI}\left(phrase, \text{“poor”}\right)\end{aligned} \tag{1.2}$$

1.2.2 国内研究现状

国内主客观文本分类研究相较国外，起步较晚。姚天昉和彭思崴[12] 利用情感形容词，人称代词，不规范的标点符号和带情感的标点符号等作为特征，比较了四类分类算法应用于主客观分类时的性能。叶强等[13] 人提出基于 2-POS 模型的中文文本主观性判断方法，分类性能接近英文类似研究结果。张博[14] 将句法关系模块，依存关系模块与 SVM 分类器结合，在实验中取得良好结果，F-measure 值达到 88.9%。

参 考 文 献

[1] MOORE G E, et al. Progress in digital integrated electronics[C]//Electron devices meeting: volume 21. [S.l.: s.n.], 1975: 11-13.

[2] 王安君, 黄凯凯, 陆黎明. 基于 Bert-Condition-CNN 的中文微博立场检测[J]. 计算机系统应用, 2019, 28(11):45-53.

[3] 周胜臣, 瞿文婷, 石英子, 等. 中文微博情感分析研究综述[J]. 计算机应用与软件, 2013, 30(3):161-164.

[4] 崔志刚. 基于电商网站商品评论数据的用户情感分析[D]. 北京: 北京交通大学, 2014.

[5] 杨立公, 朱俭, 汤世平. 文本情感分析综述[J]. 计算机应用, 2013, 33(06):1574-1607.

[6] 杨小平, 张中夏, 王良, 等. 基于 Word2Vec 的情感词典自动构建与优化[J]. 计算机科学, 2017, 44(1):42-47.

[7] 朱晓霞, 宋嘉欣, 张晓缇. 基于主题挖掘技术的文本情感分析综述[J]. 情报理论与实践, 2019, 42(11):156-163.

[8] WIEBE J, BRUCE R, O'HARA T P. Development and use of a gold-standard data set for subjectivity classifications[C]//Proceedings of the 37th annual meeting of the Association for Computational Linguistics. [S.l.: s.n.], 1999: 246-253.

[9] RILOFF E, WIEBE J. Learning extraction patterns for subjective expressions[C]// Proceedings of the 2003 conference on Empirical methods in natural language processing. [S.l.: s.n.], 2003: 105-112.

[10] WILSON T, WIEBE J, HOFFMANN P. Recognizing contextual polarity in phrase-level sentiment analysis[C]//Proceedings of human language technology conference and conference on empirical methods in natural language processing. [S.l.: s.n.], 2005: 347-354.

[11] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C/OL]//ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. USA: Association for Computational Linguistics, 2002: 417‑424. https://doi.org/10.3115/1073083.1073153.

[12] 姚天昉, 彭思崴. 汉语主客观文本分类方法的研究[C]//第三届全国信息检索与内容安全学术会议论文集. 苏州: 中国中文信息学会, 2007: 117-123.

[13] 叶强, 张紫琼, 罗振雄. 面问互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007, 1(1):79-91.

[14] 张博. 基于 SVM 的中文观点句抽取[D]. 北京: 北京邮电大学, 2011.