Statistics and Application 统计学与应用, 2023, 12(2), 406-410
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122044

Hans汉斯

基于文本挖掘技术校园学生投诉问题分析

朱美瑶，涂现峰，王宇喆，钟美君

嘉兴南湖学院，信息工程学院，浙江 嘉兴

收稿日期：2023年3月24日；录用日期：2023年4月14日；发布日期：2023年4月26日

摘　要

本研究主要探索从文本中获得有效信息的问题。本文以投诉学校的文本为处理对象，调用Python对投诉文本分析、词频统计以及主题分析，从而了解投诉文本的集中问题点，为快速有效解决学生投诉问题、合理处理学校与学生的关系提供参考，同时指出对投诉问题分类是值得深入探讨的问题。

关键词

文本挖掘，投诉，主题分析，Linear Discriminant Analysis (LDA)

Analysis of Student Complaints Based on Text Mining Technology

Meiyao Zhu, Xianfeng Tu, Yuzhe Wang, Meijun Zhong

School of Information Engineering, Jiaxing Nanhu University, Jiaxing Zhejiang

Received: Mar. 24th, 2023; accepted: Apr. 14th, 2023; published: Apr. 26th, 2023

Abstract

This study focuses on exploring the problem of obtaining effective information from texts. In this paper, we take the text of complaints against schools as the processing object, and use Python to analyze the complaint text, word frequency statistics and topic analysis, so as to understand the concentrated problem points of the complaint text, and provide reference for quickly and effectively solving student complaints and reasonably handling the relationship between schools and students, and also point out that classifying the complaint problems is a problem worthy of in-depth exploration.

Keywords

Text Mining, Complaints, Topic Analysis, Linear Discriminant Analysis (LDA)

文章引用: 朱美瑶, 涂现峰, 王宇喆, 钟美君. 基于文本挖掘技术校园学生投诉问题分析[J]. 统计学与应用, 2023, 12(2): 406-410. DOI: 10.12677/sa.2023.122044

1. 引言

随着时代的发展，数据收集技术和数据存储技术快速进步，这也让各个场景可以积累大量而富有特点的数据，数据中隐藏价值，数据挖掘将传统分析方法与复杂算法结合，为各领域数据的处理提供了机会。对文本数据而形成的文本挖掘方法也在实际应用中发挥着重大作用，比如商家收集用户对产品的评论数据，运用文本分析技术分析用户的喜好、关注点，提升产品的质量或推送合适的相关产品。

2. 研究背景

2.1. 文本分析应用普及性

文本分析是指从文本中获得有价值信息的方法，文本分析作为数据挖掘中的一种方法有多重用途，如运用文本分析技术从知网 CSSCI 数据库中挖掘关键词的先验信息，以此设计问卷更精准的研究乡村产业振兴的金融溢出效应及其影响因素[1]；如通过收集微博数据，提出训练模型从微博文本中抽取出该文本是否包含定义的事件、事件类型、元素等信息，从而探究用户个人时间轴上的事件变化规律来预测个人事件，以实现对用户的个性化推荐[2]；如轨道电路故障文本利用率低且比较依赖人工分析，运用文本挖掘技术，从故障致因出发实现故障文本特征表示，进而实现故障文本自动分类，对现场检修、预防性维护给出指导[3]；如利用文本挖掘对大量的图书信息进行筛选，模拟用户的使用习惯，提高图书信息管理效率，降低维护成本等[4]。可以看出，文本挖掘不仅可以在各个场景下产生实际有价值的作用，还可以用在研究工作前的探讨，可以说在目前网络的发展下，文本数据蕴含着极大的价值，但文本中书写不规范、口语化等一些对分析不利的特点，如何更好地利用文本分析仍是一个不断研究的热点问题。

2.2. 校园投诉处理的重要性及处理现状

校园投诉事件折射着社会转型期日益凸显的教育冲突，是不同主体沟通变异，以及不同站位的利益考量产生的内在矛盾[5]。学生投诉是每个学校都会遇到的问题，是学生真实反应自己在生活学习过程中遇到的问题。就教学投诉而言，若建立完善的教学投诉处理流程，可增强教师与学生的凝聚力，促进教学良性发展[6]。近年来，时常发生因不及时处理学生投诉，导致出现学生自残、自杀、群体性事件等情况，给学生身心健康、学校正常教学和社会安全稳定带来了严重影响。只有认真对待学生投诉，快速分析，准确回应，才能有效提升学生对学校的整体满意度，及时消除问题隐患，提升校园治理能力和水平。

目前，所有的投诉都处在人工分类阶段，而且投诉数据有多种来源，如班主任、辅导员、职能部门甚至网络吐槽、信访等，最终都会以非结构化的文本信息转达，如何对这些学生的投诉信息进行有效、快速、准确的分类成为处理学生投诉、提升学生满意度的关键。

3. 研究过程

3.1. 研究思路

文本分析技术现已初具完备性，无论是模型算法层面，还是实现工具层面，都已得到有效发展，这为文本分析的研究提供了技术保证；越来越多的群体喜欢在网络上发表自己的看法，或吐槽或赞美等，这些数据的获取也具有开放性，为文本分析的研究提供了数据准备；校园投诉问题的分析不具有整体性，

一般是人工逐条处理，文本整体的关注类别，所隐含的主题这些问题的解决需以整体的视角考察，基于以上几点本研究对校园投诉问题运用文本分析技术挖掘主题，以确定投诉问题的关注要点，为更好地解决校园投诉问题提供参考。

3.2. 研究数据

数据来源于百度贴吧中的投诉学校吧，共收集 129 条文本数据，通过挖掘这些文本信息，以便更好地解决学生投诉问题，文本数据示例如表 1。

Table 1. Text examples of student complaints
表 1. 学生投诉文本示例

序号	回答
1	劝看到了就别去这个学校了，学生偷东西当小偷，被偷价值好几千的东西，然后还多次偷取，学校是干嘛的，育人，教人偷东西的？如果不想让这个锅，让学校背着，只有找出小偷，不然，这个学校去了也是让学生当小偷的，有什么用，倒不如不要花那么多钱，去上一个这样的学校，差劲！
2	怎么样投诉学校对即将要实习(还有两个月)的大学生搬宿舍?因为今年招的学生不多，学校就想把我们这栋的空出来，我们这七八个宿舍人搬到其他栋，我们有没有权利要求不搬呢？
3	这个学校乱收高费，一个学期就收一百多元的费用，不知道是用做什么的，具体不明。
4	投诉外语外贸中招 1 班的平面设计老师，还是师范的研究生，上课 7 什么都不教直接拿图给你画，我画成这个样子他说我没用颜色涂色直接撕了说没做作业扣学分你们说这种老师还有用吗？
5	这个学校动不动就拿开除学籍来威胁我们学生
...	...

3.3. 研究过程及结果

3.3.1. 分词

文本数据准备后，接下来对文本内容进行分词处理，即将一段文本分成相互的独立的词，中文文本的话 jieba 分词相对比其他工具有较好的效果，故本研究通过 Python 中的 jieba 库对文本数据分词处理，文本中有很多对分析无用的字符或一些语气助词，如“的”、“在”等等，分词后利用 Python 去停用词。

3.3.2. 词频分析

对投诉文本中词汇出现的次数进行统计和分析，通过词汇的频次变化，初步分析投诉文本的关注点，投诉文本中排名前 15 次的词频如表 2，可以看出：① 学校、学生、老师排在前列，这三者都是学校主体中涵盖的基本要素，说明这三者关系的处理是解决投诉问题的核心；② 补课、上课等词排在前列，这是学生在校园内的主要活动，说明合理安排上课是解决投诉问题的关键；③ 家长、孩子等词高频出现，说明校园投诉问题的解决家庭也是关键一环，文本数据示例如表 2。

Table 2. Top 15 high-frequency words of complaint text
表 2. 投诉文本前 15 高频词汇

排序	词汇	频数	排序	词汇	频数	排序	词汇	频数
1	学校	151	6	中学	19	11	吐槽	13
2	学生	81	7	补课	16	12	核酸	12
3	老师	44	8	班主任	14	13	校长	11

Continued

4	孩子	37	9	投诉	13	14	真的	11
5	家长	24	10	上课	13	15	领导	11

3.3.3. 主题分析

LDA 主题模型是计算机化文本分析的重要方法之一，本部分运用主题模型对投诉文本进行探索性研究，以达到对文本信息更深层次性的挖掘。运用 LDA 主题模型对评论文本进行主题挖掘前，需要先确定主题的数量，主题过少无法解决主题的内容细节，主题过多模型会过拟合，因此模型确定合适的主题数量很重要，困惑度是用来评价语言模型好坏的指标，可用于确定主题数量[7]，如图 1，随着主题数目的增加，困惑度不断下降，主题数目为 4 后，困惑度下降的速度变缓，故评论文本的主题数目设置为 4，文本数据示例如图 1。

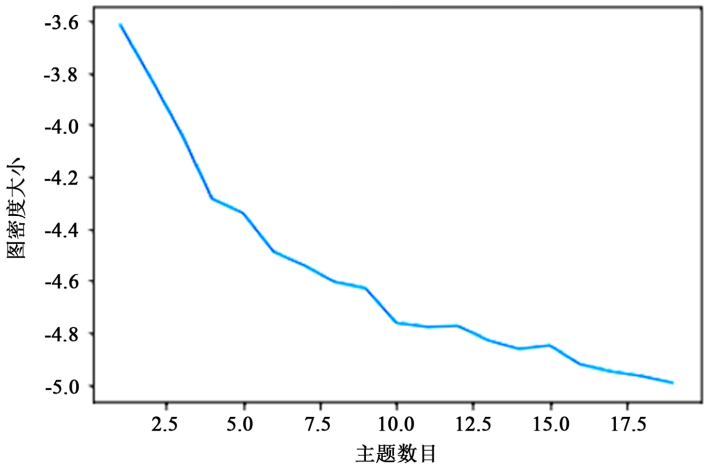

Figure 1. Subject of complaint text—change in confusion
图 1. 投诉文本主题——困惑度变化情况

模型可视化结果各个主题间无交叉，模型效果较好。投诉文本的主题与其对应的词汇如表 3，有 4 个主题，分别为学校主体元素、学校休息安排、学生学习、学生管理团队，可以看出投诉文本体现在学校与学生的关系、学生的学习生活上，学校主体元素应更多考虑到学生的学习生活的便利性，学生的学习生活应留一定的自主支配时间，学校管理团队应更多从学生角度考虑问题，文本数据示例如表 3。

Table 3. Subject extraction results of complaint text
表 3. 投诉文本主题提取结果

主题	特征词
Topic 0 (学校主体元素)	学校、老师、上课、宿舍
Topic 1 (学校休息安排)	假期、学习、学期、工作、寒假
Topic 2 (学生学习)	学生、补课、学习、一点、放假、专业课
Topic 3 (学校管理团队)	校长、领导、班主任、安排

4. 结语

文本挖掘已应用到现实中的很多场景中，但在校园社会治理信息服务上利用好文本信息仍需要进一

步探究。本研究主要探讨投诉学校文本的主题情况，但利用模型为平台用户的投诉进行自动分类，精准推送相关服务部门，提升校园治理系统化、精细化、智能化水平，形成与高质量发展相适应的校园治理体系是值得继续深入解决的问题。

参考文献

[1] 刘赛红, 孙媛. 乡村产业振兴的金融溢出效应及其影响因素研究——基于文本挖掘与问卷调查[J]. 经济问题, 2022(12): 53-62.

[2] 肖锐, 刘明义, 涂志莹, 王忠杰. 基于社交媒体文本挖掘的个人事件检测方法[J]. 计算机应用, 2022, 42(11): 3513-3519.

[3] 侯通, 郑启明, 姚新文, 陈光武, 王小敏. 基于文本挖掘的轨道电路细粒度故障致因分析方法[J]. 铁道学报, 2022, 44(10): 73-81.

[4] 余贤锋, 杨新彬, 张文岳, 司占军. 基于文本挖掘与推荐系统的图书管理软件(英文) [J]. 数字印刷, 2022(5): 58-63.

[5] 蔡辰梅, 席长华. 学校缘何被投诉?——转型社会中的多元主体教育冲突案例研究[J]. 全球教育展望, 2022, 51(9): 23-39.

[6] 王智松. 教学投诉蕴藏的价值与“教学投诉处理流程”的建立[J]. 高等继续教育学报, 2013, 26(6): 49-51.

[7] Huang, L., Ma, J.Y. and Chen, C.L. (2017) Topic Detection from Microblogs Using T-LDA and Perplexity. *Proceedings of* 2017 24*th Asia-Pacific Software Engineering Conference Workshops* (APSECW), Nanjing, China, 4-8 December 2017, 71-77. https://doi.org/10.1109/APSECW.2017.11