

基于朴素贝叶斯模型的一种网络负面信息预警策略研究^{*}

张 扬¹ 崔晨阳²

(1 中国人民公安大学 2 北京大学)

摘 要 Naive Bayes是一种基于概率的分类器,它用各个类别的先验概率和每个类别出现特定特征的条件概率来预测出现这些特征的个体的类别。针对当前“网络负面信息满天飞”的现状,本文提出了一种基于朴素贝叶斯模型的网络负面信息预警策略。与一般的文本分类不同,针对大规模网络碎片化信息的情感识别一方面对执行效率要求很高,另一方面主要关注有主观情感倾向的词。针对这些问题,我们做了相应的优化策略,如提取情感倾向专用停用词表,细化对否定词的处理等,并以2万条微博数据样本为例进行测试,实验证明这些策略在文本情感识别中具有较为理想的执行效率和准确率。

关键词 负面信息 情感分析 机器学习 朴素贝叶斯 舆情监测 预警

DOI: 10.13663/j.cnki.lj.2014.08.012

Research on Early Warning Strategy of Negative Information on the Internet Based on Naive Bayes Model

Zhang Yang¹, Cui Chenyang²

(1 People's Public Security University of China, 2 Peking University)

Abstract Naive Bayes is a kind of classifier based on probability and used for the prediction of individual categories with the prior probability and conditional probability of each category. Targeting at the current “negative information all over the Internet” phenomenon, the paper offers an early warning model based on Naive Bayes method. Different from general text classification, emotion recognition aimed at large-scale network information mainly focuses on words with subjective emotiosn and requires high execution efficiency. To solve these problems, we conducted the corresponding optimization, such as extracting Emotional Tendency Stop Words List, detailing the management of negative words, and taking 20000 Twitters as sample to test the effectiveness of the model on text emotion recognition. Experiments showed these strategies have ideal execution efficiency and accuracy.

Key words Negative information, Sentiment analysis, Machine learning, Naive Bayes, Public opinion monitoring, Early warning

随着网络新媒体的快速发展和迅速普及,大多数人已经习惯从“百度”获取资讯,通过“微信”传递动态,使用“微博”记录感想。然而,网络作为一块自由领地,其信息良莠不齐,鱼龙混杂。特别是2011年以来,伴随微博、微信、移动新闻客户端的爆发式增长,色情、暴力、迷信、极端、贪腐和一些带有煽动、挑拨、污蔑、抹黑倾向的负面新闻以较低门槛涌入网络,在起到媒体监督等积极作用的

同时,也极易使网民“个体的有意识”转变为“群体的无意识”,从而导致群体极化和自我否定,积聚社会负面情绪^[1]。因此,要将负面新闻的负效果降到最低程度,必须对其进行有效监控和预警,从而有针对性地开展正面引导。

^{*} 本文系中国人民公安大学研究生创新项目“基于模式匹配和机器学习的网络舆情情感倾向性分析模型研究”(项目编号:2013SKX04-5)的研究成果之一。

1 负面信息的内涵及特点

负面信息是与正面信息相对的一个概念，泛指有害的、错误的、低俗的、消极的新闻、报道、评论等信息。在网络日益成为社会思潮、社会舆论重要策源地和集散地的形势下，负面信息在一定程度上对网民的情绪、价值观、思想道德等都会产生潜移默化的影响。随着社会向多元化发展，人们只能借助网络去了解复杂的大千世界^[2]。因此，人们对社会的认知恰恰是对网络环境的反映^[3]。然而，网络上的负面信息多是“天灾人祸、违法犯罪、贪污腐败、变态色情”等破坏性的新闻事实，往往会通过华美鲜艳的外包装和“不容置喙”的逻辑链，对普通民众产生“润物细无声式”的腐蚀，极易使人们对现实环境产生抵触、怀疑，对社会失去信任。例如广为人知的富士康跳楼事件和校园暴力事件，虽然各大网站只是客观报道事实真相，但是短时间内的大量负面报道影响了人们对真实环境的认知，反而助推了极端恶性事件的发生。

一直以来，我国部分地方政府和相关部门对负面信息“讳莫如深”，往往采取网上屏蔽与删除、网下专项治理和运动式干预等手段来封堵负面信息源，效果很不理想。建立高效、智能的网络负面信息识别预警系统，及时把握和引导负面信息的传播是非常必要的和紧迫的。

2 系统设计

2.1 系统概况

本系统设计目标是处理网络上海量分散信息，诸如微博、微信、新闻、评论等，以快速发现、提取、推送负面信息、评论等。因此系统的执行速度是很重要的因素，在此我们采用效率较高的朴素贝叶斯方法作为基本的分类方法，朴素贝叶斯方法的优缺点以及其他相关技术工作的对比将在 2.2 节给出。3.2 节的实验结果表明，本系统在具备较快的执行速度的前提下，给出了正确率较高的识别结果。其速度和正确率能够满足预设的目标。

本系统分为训练模块与情感分析模块两个模块。训练模块输入已标注好的数据，输出训练好的模型。情感分析模块输入未标注的数

据，输出对该数据的预测结果。所有的数据都是以行为单位：

数据内容 {1, 0}

数据内容表示预测的文字内容，1 表示正面情感，0 表示负面情感，中间以制表符隔开。

标注和预测的结果仅分为正、负两类。情感分析模块的数据结果会呈献给用户，用户可以进行修正，修正后的结果又会被作为训练模块的输入数据。随着使用量的增加，训练数据也会不断地增加，系统将持续进行训练过程和情感分析过程。3.2 节的实验结果表明，随着训练数据量的增加，模型的预测能力会不断改进，训练数据量到 1.2 万条后会产生比较理想的预测结果。

系统运行界面如下图所示：

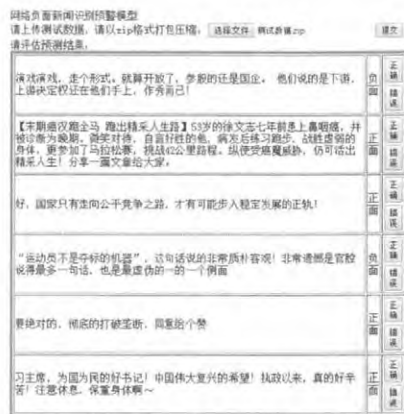


图1 网络负面信息识别预警系统运行界面

2.2 相关研究

文本信息的情感识别多使用基于主题和基于情感两种方法。二者使用的基本分类方法是一致的，区别仅仅体现在选取的特征上。所以一些关于基于主题的分类的相关经验可以借鉴到我们的研究上。

对于一篇文章而言，其语义倾向性主要取决于其所用词汇的倾向性，不同的词语有不同的倾向性，确定这些词汇的倾向性问题就成为本文研究的基础。然而，但是用静态的方法来统计所有在网络中出现的词语的情感倾向性是一件不可能的事情。这之所以是一件不可能的工作，一方面因为词语的数量巨大很难统计完全，另一方面每个词在不同语境中的倾向性也是不断变化的。历史上有些人利用一些 NLP 的

方法来推测词语的倾向性,如 Hatzivassiloglou^[4]等人,他们所依据的是词语之间的关系。他们的工作基于这么一个假设:如果形容词之间有“and”、“but”之类的连接词,那么我们很容易可以从一个词的倾向性推断出另一个词的倾向性,因为它们只能是简单的传递或取反。例如文中出现了“excellent and X”,根据之前的假设,我们可以断定 X 也含有正面倾向。基于该假设,Hatzivassiloglou 创造了 4-step 算法。也有人如 Turney^[5]等利用一个预先定义的具有明显倾向性的种子词集合,用统计的方法计算其他词语与其相关度来判断目标词语的倾向性,他们提出了两种方法:PMI2IR 和 LSA。Esuli^[6]利用对词语外部信息的学习分类判断语义倾向,主要依托词典或者注释中的信息。

国内外有关中文词句情感分析的研究相对较少。Yuen^[7]统计出了一组具有强烈语义倾向的词汇表,然后统计文中每个词语与词汇表中的词语的统计关系,借此估计每个词的语义倾向。朱嫣岚等^[8]使用了一种词汇语义网来进行语义相似度比较,从而进行词语的语义倾向性预测。

这些研究大多数是根据褒义词或者贬义词的统计分布来进行学习分类,从而预测句子的倾向性。例如 Turney 在用前面提到的 PMI2IR 方法计算出符合预设规则的短语的语义倾向的基础上,通过对文章中所有词语的倾向性指数求平均值来确定整个文章的倾向性。通过 Pang Bo 的研究可知,在应用朴素贝叶斯模型、最大熵模型和 SVM 对电影评论进行分类的结果中,SVM 的效果最好,准确率高达 80%,但是 SVM 计算时间复杂度非常高,即使用 SMO 算法,所需时间仍然是朴素贝叶斯模型的 5 倍。而且 SVM 模型对特征的提取要求很严格,其分类效果在很大程度上取决于特征的选择,不同的分类目标,不同的语言,不同的测试语料所需的特征都不相同。对于微博数据来说,其数据格式灵活,内容多变,很难找到一组通用的特征,选择的特征很容易就会产生过拟合。所以我们认为,SVM 模型不适合微博数据等一些内容差异性很大的网络数据。

2.3 分类方法

由于我们要处理的数据是大量分散的互联网碎片化的信息,处理的效率问题就至关重要,

因此我们采用朴素贝叶斯模型来进行文本分类。文本用向量空间模型表示, $\{f_1, f_2, f_3, \dots, f_m\}$ 表示文章中能出现的所有的 m 个特征,用 n_i 表示 f_i 的权重,这样一篇文章就可以表示为一个特征序列 $d = \{n_1, n_2, n_3, \dots, n_m\}$ 。其中 N_j 表示满足条件的文档个数, N 为总样本数, n_i, j 表示特征 i 在文档 j 中出现了多少次。

2.4 naive bayes

naive bayes 是一种基于概率的分类器,它用各个类别的先验概率和每个类别出现特定特征的条件概率来预测出现这些特征的个体的类别。在计算条件概率时引用了朴素贝叶斯假设,即每个特征出现的概率是独立的。之所以引入朴素贝叶斯假设是为了避免特征组合过多,以至于数据过于稀疏的问题。朴素贝叶斯的计算用数学表示:

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)} \\ = \frac{P(c_j) \prod_{i=1}^m P(f_i|c_j)^{n_i}}{P(d)}$$

对于我们的工作来说,上式的分母是不变的,由于我们只关注相对大小,所以可以只计算分子部分。我们赋予文本概率最大的类别。经过多次训练后,得到 $P(c_j)$ 和 $P(f_i|c_j)$ 的估计:

$$\hat{P}(c_j) = N_j/N \\ \hat{P}(f_i|c_j) = \frac{1 + n_{i,j}}{m + \sum_{k=1}^m n_{k,j}}$$

朴素贝叶斯分类器最根本的特点是引入朴素贝叶斯假设,即文档中的词是条件独立的,这是一个很强的假设,虽然不很符合常识但却效果很好。具体训练和预测算法如下图所示:

```

TRAINMULTINOMIALNB(C,D)
1  V ← EXTRACTVOCABULARY(D)
2  N ← COUNTDOCS(D)
3  for each c ∈ C
4  do Nc ← COUNTDOCSINCLASS(D,c)
5  prior[c] ← Nc/N
6  textc ← CONCATENATETEXTOFFALLDOCSINCLASS(D,c)
7  for each t ∈ V
8  do Tct ← COUNTTOKENSOFTERM(textc,t)
9  for each t ∈ V
10 do condprob[t][c] ←  $\frac{T_{ct}+1}{\sum_{c'}(T_{ct'}+1)}$ 
11 return V, prior, condprob

APPLYMULTINOMIALNB(C,V,prior,condprob,d)
1  W ← EXTRACTTOKENSFROMDOC(V,d)
2  for each c ∈ C
3  do score[c] ← log prior[c]
4  for each t ∈ W
5  do score[c] += log condprob[t][c]
6  return arg maxc∈C score[c]

```

图2 朴素贝叶斯训练和预测算法

2.5 模型优化

自然语言处理领域的很多工作都集中在数据的前期处理上，这类工作十分繁琐，但对结果的影响至关重要。我们在做这个预警模型时也做了大量的前期数据处理工作，这些工作对最终的准确率有着非常大的影响。

我们发现，一些通用的停用词表在我们的系统下表现并不十分的理想。这主要是因为，这些通用停用词表是针对基于主题的各种语言模型提取出的，提取出的一些常用词对主题的贡献不大，但是并不代表这些词对情感倾向的贡献也很小。比如，“好”，“的确”被收录到很多通用性的停用词表中，但这些词实际上是对表达正面情感倾向贡献很大的词。基于我们需求和现有条件的矛盾，我们用 topic model 的变形 topic in set 抽取出一一些“白名单”词，使其保证不出现在停用词表中。

对否定词的处理也非常的重要，因为否定词之后的情感倾向完全和原义相反，这会对结果造成很大的干扰。Sanjiv^[9]在采用了一种简单的方法处理这个问题，他对从否定词到首次出现的标点符号中间的所有词语标记上“_N”后缀，以此来处理否定词对一句话的语义倾向的影响。但我们实验发现这种方法引入了过多的无关项，只很小幅度地提高了准确率。通过研究网络上的语料发现，否定词否定的对象主要是该句话中所有的动词、名词和形容词。本文采用的处理方法是将这三种否定对象后加上“_N”后缀。

在第3节的实验设计和分析中，我们分别作了不同的实验对我们的优化策略进行检测，结果证明，我们的优化策略在很大程度上提升了预测准确率。

3 实验结果与分析

本次试验使用的数据是从新浪微博上抓取的2万条微博内容。经过人工标注倾向性，从这2万条数据抽出3千条做测试数据，1.7万条做训练数据。人工标注的只有两类，正类和负类，正类表示有负面情感倾向的内容，负类表示未有负面情感倾向的内容。

3.1 实验设计

在处理每条数据的时候，一步必须要做的

工作是去除停用词，目前有很多停用词表供我们使用，比如哈尔滨工程大学停用词表，四川大学职能科学实验室的停用词表等。然而我们研究的问题有一定的特殊性，我们做的是基于情感的分类问题。很多对我们工作有明显影响的词可能会被收录在停用词表中，而一些对我们影响不大的词却没被收录。所以针对我们的特定问题，特整理出情感识别分类专用停用词表。实验一就是为了验证不同的停用词表对于情感识别效果的影响。实验中，我们分别同样的训练语料，对其进行三种不同的预处理，即分别用哈尔滨工程大学停用词表、四川大学职能科学实验室停用词表和我们的专用停用词表进行过滤。然后再将过滤后的语料转换为向量空间模型的表示形式，在进行朴素贝叶斯模型的训练，最后用测试数据测试，统计准确率。

由于否定词极易导致整个语意环境的变化，实验中，我们拒绝把否定词当做停用词进行删除，而是专门设计了实验二，以研究否定词的处理与文本情感分类效果的关系。实验中，我们设置两组实验组，一组对训练语料做上述的否定词预处理，一组不做。然后在做相应的训练、测试、统计、对比工作。

另外，系统的冷启动问题也是机器学习方法经常遇到的一个难题，即如何在没有用户使用的情况下，保证第一次预测结果的准确性^[10]。针对我们的工作，这个问题的一个可行的解决方法就是预先用一部分标注好的数据训练出一个准确率比较高的模型。实验三就是为了测试多大是数据量能训练出比较理想的初始模型。在实验中，我们分别用不同数据量的训练数据训练模型，再用同一组测试数据测试准确率。

3.2 实验结果

实验一、二、三的结果分别见表1、表2和图3：

表 1 实验结果（实验一）

	哈工大停用词表	四川大学停用词表	情感识别停用词表
准确率	62.3%	65.7%	79.8%

表 2 实验结果（实验二）

	不作处理	Sanjiv 方法	本文方法
准确率	64.3%	52.2%	72.3%

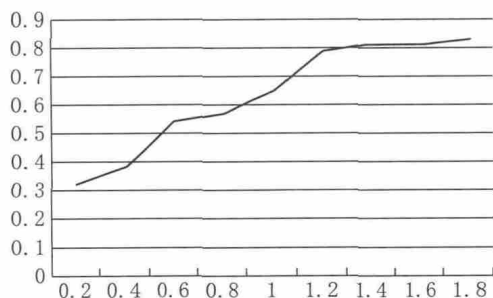


图3 实验结果(实验三)

实验一的结果表明,使用我们整理的停用词表能较大幅度的提升准确率,证明了情感识别领域的停用词和其他领域有较大区别这一假设。

实验二的结果表明,对否定词不加任何处理时准确率较低,而用 Sanjiv 方法进行处理后的准确率反而更低。这可能是因为这种方法对太多的词进行的特殊化,而情感识别又对这些词比较敏感,因此会产生负面影响。本文方法效果最好,由此证明上文的假设是正确的。

参考文献

- [1] 埃里克·霍弗. 狂热分子:群众运动圣经[M]. 第2版. 梁永安, 译. 桂林: 广西师范大学出版社, 2011:20-23.
- [2] 李普曼. 舆论学[M]. 林珊, 译. 北京: 华夏出版社, 1989:50-52.
- [3] 郑保卫. 当代新闻理论[M]. 北京: 新华出版社, 2003:5-7.
- [4] Vasileios Hatzivassiloglou, Kathleen R. McKeown. Predicting the Semantic Orientation of Adjectives [C]//Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, 1997:174-181.
- [5] Turney Peter. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 417-424.
- [6] Esuli, Andrea, Sebastiani, Fabrizio. Determining the Semantic Orientation of Terms Through Gloss Classification [C]//Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management, 2005: 617-624.

实验三的结果表明,当训练数据量达到1.2万条以上时,准确率会稳定在较高的水平。这个结果对我们系统的前期初始模型训练有较大的指导意义。

4 小结

通过运用朴素贝叶斯方法,针对网络信息的特点,提取情感倾向专用停用词表,细化对否定词的处理,优化情感识别策略,创新舆论监测模型,使用机器学习方法来实现基于情感的文本分类,可以对网络负面信息进行有效监控。本文构建的网络负面信息识别预警模型,操作简便,分析准确,响应及时,且根据性能测试的实验数据,得知在实际的运行环境下具有较好的执行效率和准确率。但是,该模型目前只是在提供特定语料的基础上进行分析,在实际应用中应当对传统的网络爬虫抓取方法进行有效改进,实时加入最新的语料,以使其具有更强的适应性。

- [7] R W M Yuen, T Y W Chan et al. Morpheme-based Derivation of Bipolar Semantic Orientation of Chinese Words [C]//Proceedings of the 20th International Conference on Computational Linguistics(COLING-2004), 2004: 1008-1014.
- [8] 朱婧岚, 闵锦, 周雅倩等. 基于How Net的词汇语义倾向计算[J]. 中文信息学报, 2005(1): 14-20.
- [9] Sanjiv D, M.Chen. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards [C]//Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.
- [10] 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007(6): 95-100.

张 扬 中国人民公安大学2012级硕士研究生。研究方向: 公安情报分析。E-mail: 15101108386@126.com 北京 100038

崔晨阳 北京大学2012级硕士研究生。研究方向: 软件与理论。 北京 100871

(收稿日期: 2014-03-03)